THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

CRISPR/Cas9-mediated traceless gene correction and activation of the *HBB* locus in iPSCs with the beta thalassaemia mutation

Suad Alateeq

BSc, MSc

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

Australian Institute for Bioengineering and Nanotechnology

ABSTRACT

Beta thalassaemia is caused by mutations in the adult haemoglobin gene (*HBB*) and is one of the most prevalent monogenic blood disorders worldwide. The transplantation of haematopoietic stem cells derived from gene-corrected patient induced pluripotent stem cells (iPSCs) could provide a promising therapeutic strategy. Here, the generation and characterisation of footprint-free iPSCs from dermal fibroblasts of an individual with a homozygous beta thalassaemia-causing mutation affecting splicing is described (Chapter III). Successful gene correction of the disease-causing mutation was achieved by employing a CRISPR/Cas9 dual nickase approach and two donor design strategies aimed at reducing undesired on-target mutagenesis (Chapter IV), followed by the seamless piggyBac transposase-aided removal of extraneous sequences. The first strategy involved the introduction of a SNP in the protospacer adjacent motif (PAM) sequence in one guide of the first gRNA pair, and the second utilised insertion of a large *piggyBac* transposase-exciseable selection cassette (>3 kb) to spatially separate the other gRNA pair. The combined targeting efficiency was ≈68.7 per cent (52 per cent heterozygous and 16.6 per cent homozygous) after excluding clones with random integration. Surprisingly, indels were observed in the untargeted allele in all of the heterozygote clones (25/25), indicating a high rate of on-target non-homologous end joining (NHEJ) driven mutagenesis. Significantly, the data also showed that the nickase variant of Cas9 elicited unexpected on-target mutagenesis also affecting the targeted allele, as insertions or deletions were observed at the binding site of one of the guides in five of the 26 targeted clones (19 monoallelic and seven biallelic targeted clones with sgRNAs 4/5). No off-target mutagenesis was detected upon screening of the five top-scored homologous genomic sites for each of the individual gRNAs (10 sites per pair). *PiggyBac* transposase-mediated excision of the selection cassette(s) demonstrated that seamless correction of homozygous clones correctly targeted at both alleles is feasible. In order to assess the restoration of normal transcription from the corrected allele, a CRISPRa-based approach to activate the *HBB* proximal promoter region was optimised and achieved a significant mRNA expression level from the corrected allele after cassette removal (Chapter V). This demonstrated that such an approach can be used to model defective splicing of the *HBB* gene in iPSCs with beta thalassaemia mutations and corroborates restoration of normal *HBB* splicing in gene-corrected cells.

This study combined cutting-edge cellular reprogramming methods to generate beta thalassaemia iPSCs, a double nickase-mediated gene editing approach and piggyBac transposase-aided excision to achieve seamless gene repair, and a CRISPRa approach to model the impact of the disease-causing mutation and demonstrate restoration of normal transcription following genetic repair. Having established a platform for precise gene correction and for validation of the restoration of the functional allele in this monogenic disease, this strategy may provide a viable avenue for iPSC-

based cell therapy of beta thalassaemic patients provided mature engraftable blood cells can be generated from hiPSCs in the future.

Declaration by the author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during the candidature

Peer-reviewed papers:

Alateeq, Suad, Fortuna, Patrick R. J. and Wolvetang, Ernst (2015) Advances in reprogramming to pluripotency. *Current Stem Cell Research and Therapy*, 10(3): 193–207.

Publications included in this thesis:

No publications included.

Contributions by others to the thesis

Ernst Wolvetang and Dmitry Ovchinnikov: Conception and design (45%), data and analysis (10%); Deanne Whitworth: Conception and design (2%), data and analysis (1%); Suad Alateeq: Conception and design (53%), Collection and assembly of data (93%), Data analysis and interpretation (89%); Timothy Tracey Collection and assembly of data (5%); Stemcore (2%).

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

ACKNOWLEDGEMENTS

The continuous support I received from supervisors, colleagues, friends and family has helped me so much throughout this intense period. It has been a period of intense learning for me, at both scientific and personal levels. I owe my deepest gratitude to Prof Ernst Wolvetang for his expert advice and feedback. His passion for science, guidance, and encouragement throughout difficult times helped me to complete this thesis. I would also like to extend my thanks to my co-supervisors, Dr Dmitry Ovchinnikov and Dr Deanne Whitworth, for the insightful comments and feedback, which inspired me to widen my research focus to include multiple perspectives.

To everyone in the stem cell engineering group who helped me throughout this period, thank you for the stimulating discussions, sharing of experience and knowledge, and more importantly, for your support. Jane Sun my wise friend, thank you for everything and for your sound advice through the difficulties in my personal life. Samuel Nayler, I have enjoyed our interesting discussions and thank you for looking after my cells while I was away. A special thank you to Timothy Tracey, I have enjoyed working with you and wish you all the best in your PhD journey. To the present SCEG team Sally, Jane, Tim, Anu, Sam, Carola, Owen, Patrick, Set yen, Nadia, Arash and Parsanna, and the past team Samah, Lindsey, Charles, Nilay, it was a pleasure working with you all.

This thesis would not have been possible without the continuous support and the funding of the of the University of Dammam especially Dr. Abdullah Al-Rubaish and Dr Amein Al-Ali. I would also like to express my sincere gratitude to Dr. Sara Al-Breiki, Dr Ahlam Al-Qatari and Dr Zaki Nasserullah for helping me with my sample collection and to Elite Editing for proofreading my thesis manuscript (restricted to Standards D and E of the Australian Standards for Editing Practice).

To my special and amazing family Iqbal, Elham, Sara, Maya, Mo, and Faisal, without whom none of my success would have been possible, I am blessed to have you all in my life. You are, and always be, my inspiration and my strength, especially my little sister, Maya, you always bring a big smile to my face, no matter what.

Lastly, this dissertation is dedicated to my late parents. Your lessons and loving memories are stored in my heart. You will always be my constant inspiration. I miss you badly.

Keywords

Induced pluripotent stem cells, paired nickase, gene correction, beta thalassaemia, gene activation, seamless correction, Cas9, Cas9D10A, beta globin, gene editing.

Australian and New Zealand Standard Research Classifications (ANZSRC)

100404 Regenerative Medicine (incl. Stem Cells and Tissue Engineering), 50%

100403 Medical Molecular Engineering of Nucleic Acids and Proteins, 40%

100401 Gene and Molecular Therapy 10%

Fields of Research (FoR) Classification

1004 Medical biotechnology : 90%

0604 Genetics: 5%

0601 Biochemistry and cell biology: 5%

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BER	Base excision repair
bp	Base pair
BT-iPSCs	Beta thalassaemia iPSCs
BT-T	Biallelically targeted iPSCs
Cas9	CRISPR-associated protein 9
Cas9DN	Cas9D10A double nickase
Cas9n	Cas9D10A nickase variant
CMV	Cytomegalovirus promoter
crRNA	CRISPR RNA
D10A	Aspartate to alanine substitution at the10th amino acid residue
dCas9	Nuclease-deficient CRISPR-associated protein 9
DDR	DNA damage response
ECM	Extracellular matrix derived from Engelbreth-Holm-Swarm murine sarcoma
ESC	Embryonic stem cells
FIAU	1-(2-deoxy-2-fluoro-beta-D-arabinofuranosyl)-5-iodouracil
GATA4	GATA binding protein 4
H840A	Histidine to alanine substitution at the 840th amino acid residue
HAP1	Human haploid cell line
HDR	Homology-directed repair
HEK293FT	Human embryonic kidney (HEK) cell lines 293FT
HMA	Heteroduplex mobility assay
HR	Homologous recombination
HSCs	Haematopoeitic stem cells
iPSCs	Induced pluripotent stem cells
K562 cells	Human immortalised myelogenous leukaemia line (erythroleukemia)
KLF4	Kruppel-like factor 4
MAP2	Microtubule-associated protein 2
MEF	Mouse embryonic fibroblast
MMR	Mismatch repair
MSCs	Mesenchymal stem cells
MT-T	Monoallelically targeted iPSCs
NANOG	Homeobox transcription factor NANOG
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining

OCT4	Octamer-binding transcription factor 4
PAGE	Polyacrylamide gel electrophoresis
Pax6	Paired box 6
pDNA	Plasmid DNA
PGK	Phosphoglycerate kinase 1 promoter
sgRNA	Single guide RNA
SMA	Smooth muscle, actin
SNP	Single nucleotide polymorphism
SOX17	SRY (sex-determining region Y) box 17
SOX2	SRY (Sex Determining Region Y)-Box 2
SV40 LT	Simian virus 40 large T antigen
Targeted allele	Allele with targeted selection cassette
TD	Transactivation domain of the virion protein (*Herpes simplex* virus) VP16
TP53	Tumour protein 53
TPs	Thomson plasmids
TRA1-60 (81)	Tumour rejection antigen1-x
tracrRNA	Transactivating crRNA
U2OS	Human aneuploid osteosarcoma cells
Untargeted allele	Original allele (with no insert) in a targeted clones
VP16	*Herpes simplex* viral protein 16
VP64	Transcriptional activator composed of four tandem copies of VP16

CHAPTER I:

LITERATURE REVIEW

1 LITERATURE REVIEW

1.1 *Introduction*

Inherited haemoglobinopathies are considered to be among the most common monogenic human disorders. Nearly 5 per cent of the world's population has been estimated to carry an allele for a 'severe' haemoglobin protein variant that can cause a disease in the homozygous state [1]. To date, allogenic haematopoietic stem cell (HSC) transplantation is the only available therapeutic approach that restores long-term haematopoietic reconstitution for some severe haemoglobinopathies, such as beta thalassaemia. This approach is hindered by the lack of availability of an immunologically closely matched donor and the possible serious complications associated with this type of therapy [2]. An alternative approach is to use autologous somatic HSCs after *ex vivo* genetic manipulation to insert a functional copy of the gene (gene addition). Although significant advances have been reported for gene addition that include the use of a safer viral vector design for inserting the therapeutic gene [3, 4], targeted repair-based approaches (non-viral) avoid the adverse effects associated with viral vectors while offering the advantage of retaining the natural context of endogenous regulatory elements for the corrected gene [5]. The targeted correction of the sickle cell disease (beta haemoglobinopathies) mutation in HSCs was recently reported [6]. Because primitive HSCs are slow-cycling (quiescent) cells and active cell division is essential for effective gene targeting [7] the approach was inefficient in primitive HSCs [5]. While this low efficiency of targeting in HSCs can in some cases be counterbalanced by the increased fitness of edited cells, as is the case with the *IL-2RG* gene [5], gene correction of beta haemoglobinopathies does not confer any survival advantage to the corrected cells. Recent advances in nuclear reprogramming and gene editing technology may be able to circumvent some of these hurdles and offer the potential for alternative therapeutic strategies [8]. However, the fidelity of genome modification remains a major concern, and limited data is available on its use in induced pluripotent stem cells (iPSCs) in the context of gene correction.

This chapter is divided into four parts. A brief overview of beta thalassaemia and methods for the derivation of iPS cell lines with beta thalassaemia mutations will be provided in the first part. This will be followed by an outline of the recent nuclease platforms used to stimulate the DNA repair mechanism, and their role in site-specific gene correction, with a focus on strategies that enhance the specificity of the Cas9 systems. The third part includes an overview of the most relevant aspects of DNA break repair mechanisms (single-strand break/double-strand break; SSB/DSB) and their role(s) in iPSCs. Lastly, the applicability of CRISPR-Cas9-mediated gene activation for the restoration of functionality of the defective β-globin gene will be discussed.

1.2 *Beta thalassaemia*

Haemoglobin is a tetrameric metalloprotein composed of a two alpha-like and two beta-like globin chains. The alpha globin cluster is located on the short arm of chromosome 16 (16p13.3), while the β-globin cluster is located on the short arm of chromosome 11 (11p15.5). The β-globin cluster contains five genes that are sequentially arranged in order of their temporal expression during ontogeny. Class switching refers to the change of globin pattern forming the tetrameric haemoglobin during development. Specifically, in the early embryonic stage, haemoglobins Grower-1($\zeta_2\varepsilon_2$), Grower-2 ($\alpha_2\varepsilon_2$) and Portland ($\zeta_2\gamma_2$) are expressed in the yolk sac (Figure 1-1.A). The globin pattern then changes to HbF ($\alpha_2\gamma_2$) as haematopoiesis shifts to the liver and bone marrow during the foetal stage. The expression of the adult phenotype (HbA: $\alpha_2\beta_2$), is delayed until postnatal development, becoming the predominant form of haemoglobin shortly after birth. Consequently, as the globin switches to the adult phenotype, the phenotype of the beta haemoglobinopathies becomes evident within the first two years in affected infants, as is the case with patients with beta thalassaemia major [9]. Reactivation of gamma globin expression has the potential to alleviate the severity of the beta haemoglobinopathy phenotype by attenuating the globin chain imbalance. Various approaches have been employed to identify modulators of HbF reactivation (class switching) [10-12], including gene editing-based studies that have potential to advance the management of severe thalassaemia [12].

Beta thalassaemia is considered one of the most common inherited disorders affecting β-globin chain synthesis in red cells[13]. It is a heterogeneous group of genetic blood disorders caused by more than 300 different mutations [14]. They are mostly inherited as an autosomal recessive trait. The most severe form of the disease is frequently associated with the homozygous or compound heterozygote state. The major determinant of the severity of the disease is the type of mutation. The β-globin gene spans 1.6 kb and encompasses three exons and two introns, in addition to the 5′ and 3′ UTRs (Figure 1-1.B). The majority of beta haemoglobinopathies are caused by point mutations, small deletions or insertions that all affect β-globin synthesis but through different molecular mechanisms: (1) RNA processing, e.g., mutations of the splice junction (IVSII-1 G>A) or in the consensus sequence flanking it (IVSI-6 T>C) that abolish or reduce the formation of the transcribed RNA product; (2) RNA translation, e.g., nonsense (CD 39 C>T) or frameshift (Fr 8/9 +G) mutations that lead to premature chain termination; (3) RNA post-transcriptional modification, e.g., changes in the polyadenylation signal (AATAAA-AACAAA); and (4) Transcriptional, e.g., mutations in the promoter element (-31 A>G) or the 5′ UTR region (+10 –T) that can also result in a decrease in β-globin chain production [14]. *HBB* gene expression is regulated by elements present

upstream of the *HBB* gene; i.e., TATA, CAAT, CACCC and by the distal long-range enhancers, known as the locus control region (LCR). The latter is defined by DNase I hypersensitive sites located 6–18 kb upstream of the ε-globin gene (Figure 1-1). Mutations in both the proximal and distal genomic regions involved in the regulation of gene expression have also been reported to affect gene expression [9]. Collectively, any of these mutations can result in either a reduction (β^+) or complete elimination (β^0) of β-globin chain synthesis. Consequently, there is an excess of unbound alpha chains that aggregate in erythroid progenitors cells, leading to ineffective erythropoiesis and causing chronic haemolytic anaemia [9]. Decreasing alpha globin chain production has been suggested as an alternative approach to improve the imbalance between alpha and beta chains in beta thalassaemia [15].

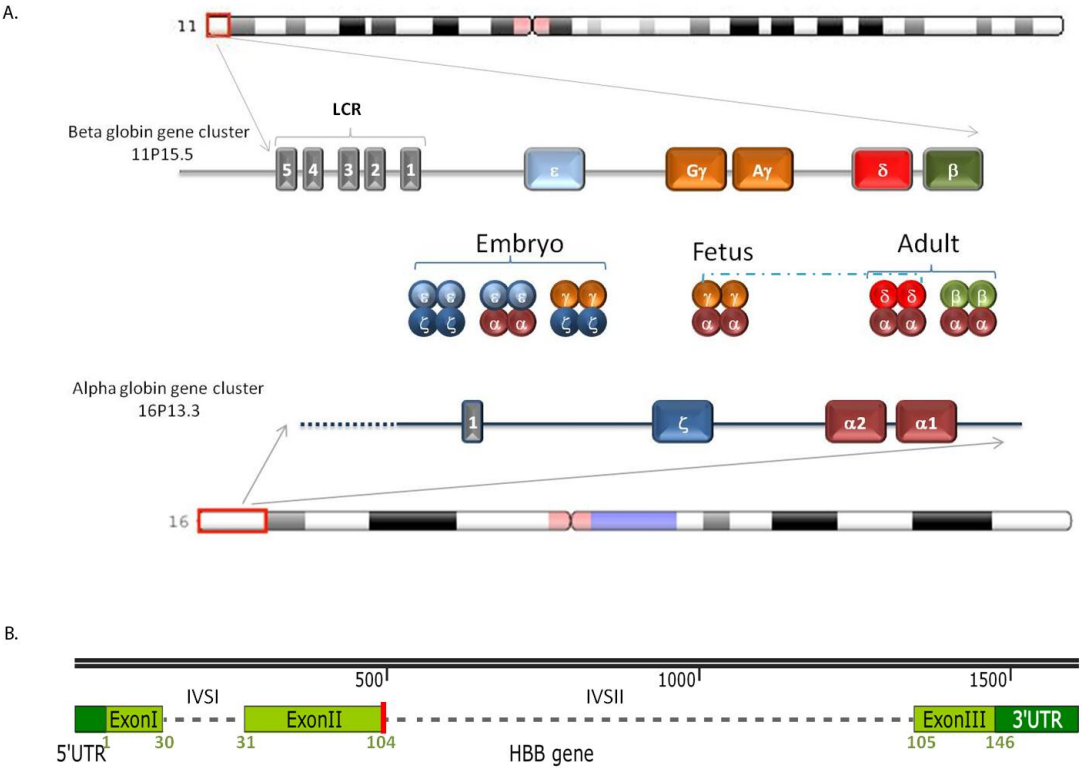

Figure 1-1: Schematic of the sequential expression of the different goblins that form the tetrameric haemoglobin during ontogeny and the structure of the β-globin gene. (A) Haemoglobins Grower-1 ($\zeta_2\varepsilon_2$), Grower-2 ($\alpha_2\varepsilon_2$) and Portland ($\zeta_2\gamma_2$) are expressed in the yolk sac, while HbF ($\alpha_2\gamma_2$) expression begins during the foetal stage and declines after birth to be replaced by the expression of the adult phenotype (HbA: $\alpha_2\beta_2$). HbD ($\alpha_2\delta_2$) and HbF form a minor component of the total haemoglobin in adult erythrocytes. Grey boxes with a number indicate distal regulatory regions. (B) Structure of the β-globin gene. The red rectangle indicates the position of the splice junction mutation IVSII-1 G>A; dark green boxes indicate the UTRs, light green boxes the exons and dashed lines the introns.

In severe forms of the disease, such as beta thalassaemia major, patients require treatment with monthly blood transfusions and iron chelation therapy. It has been estimated that approximately

40,000 infants are born with beta thalassaemia each year [1], of whom more than half will be transfusion dependent. Allogenic bone marrow (HSC) transplantation is the only therapeutic approach currently available for these patients. However, a major limitation for the majority of affected patients is finding an immune compatible donor and the serious long-term complications associated with this type of treatment [2].

1.3 *iPSCs and beta haemoglobinopathies*

In 2007, Takahashi and colleagues demonstrated that human somatic cells can be reprogrammed to produce iPSCs [16]. Nuclear reprogramming introduced the possibility of engineering cell identity and reverting terminally differentiated cells into a state, termed iPSCs, that are functionally equivalent to embryonic stem cells [16]. This was achieved through the ectopic expression of a handful of transcription factors that enabled the manipulation of cell identity. Two different combinations of the reprogramming factors OSNL (*OCT4*, *SOX2*, *NANOG*, and *LIN28*), or OSKM (*OCT3/4*, *SOX2*, *KLF4*, and *c-Myc*) were initially used to facilitate the conversion of human adult somatic cells into iPSCs using retroviral vectors [16, 17]. The approach was further optimised to generate cells with minimal impact on their genomic integrity [18].

iPSCs provide a powerful tool that can be used to generate cell models that facilitate the exploration of disease mechanisms and the study of the functional relationships between genotype and phenotypic outcomes in haematological disease [19]. They can also be differentiated into various types of patient-specific cells, including HSCs [20, 21]. iPSCs thus hold great promise for isogenic cell-based therapy, where a patient's specific somatic cells can be reprogrammed to the pluripotent state, genetically corrected, differentiated into haematopoietic progenitors with subsequent potential to be useful as a patient-specific cellular therapy [22]. Proof of concept has been realised in a mouse model with sickle cell anaemia, where the genetic defect was corrected in iPSCs and used to derive autologous HSCs. Transplantation of the corrected HSCs resulted in an improvement in the pathological phenotype [23]. Combining iPSC generation and genetic engineering technologies provides a platform for the testing of novel therapeutic drug strategies, disease modelling of genetic variants, and offers great potential for generating autologous, gene-corrected HSCs that may be suitable for possible future clinical application [24, 25].

Nonetheless, important aspects of both iPSCs and gene editing technologies remain to be addressed. Maintaining the genomic integrity of cells and establishing a balance between efficiency and specificity in gene editing are still under evaluation. Optimising technologies that have minimum impact on the host genome is therefore an active area of research.

1.3.1 Methods for deriving iPSCs with a beta thalassaemia mutation

The method of iPSC derivation can affect the quality of iPSCs, both in terms of their genome integrity, differentiation potential and compatibility with clinical applications. Integrative approaches, via lentiviral and retroviral vectors, have been used in several studies to reprogramme somatic cells from beta thalassaemia patients into iPSCs [26-28]. The main advantage of viral-based reprogramming is the high efficiency of the system compared to most non-integrative approaches [18]. However, integration-based strategies are not suitable for the production of high quality iPSCs. Indeed, integrative vectors have been shown: 1) to be associated with a high risk of tumorigenic insertional mutations that may result in undesired phenotypic alterations [29]; 2) to exhibit incomplete silencing of the vector-derived transgenes or inappropriate reactivation of the transgenes upon differentiation, thus influencing their differentiation potential [30, 31] and 3) may affect the immunogenicity of iPSCs [32, 33]. Integration-free methods can avoid these problems and offer a safer alternative for mediated cell reprogramming. To date, various methods for the generation of integration-free iPSCs have been reported in the literature [18, 34].

The Sendai virus [35] excisable integrating vector (STEMCCA/Cre/loxP) [36] and episomes [37, 38] were used for the derivation of integration-free iPSCs from patients with inherited haemoglobinopathies (Table1-1). Although transgene-free iPSCs can be derived with the STEMCCA vector, the excision process leaves a short DNA sequence in the host genome [39] that may have consequences for the functionality of the gene [40]. Footprint-free iPSCs can be derived using episomal, Sendai virus, mRNA or protein-based approaches. However, protein and mRNA-based approaches are technically laborious, as these require multiple rounds of transfection [41-43], making them less suitable for scale-up in clinical settings. In contrast, Sendai-based reprogramming offers the advantage of an efficient process, but is costly and the transgene can persist for up to 20 passages in iPSCs *(*unpublished data). Episomal reprogramming, on the other hand, is cheap and results in rapid loss of the vector during passive cell division [44]. The EBNA1 (Epstein-Barr nuclear antigen-1)/oriP-based episomal vectors have been successfully used for the derivation of integration-free human iPSCs from multiple cell types, including human fibroblasts [45, 46], cord blood mononuclear cells [47] and peripheral blood mononuclear cells [48]. In comparison to a standard plasmid, EBNA1-containing vectors have the advantage of ectopic, autonomous replication in synchrony with the host chromosome [49]. In this study, this system was chosen to generate footprint-free iPSCs with a beta thalassaemia mutation. Derivation of iPSCs were attempted from the fibroblasts of patients homozygous for the frequent Mediterranean splice junction mutation IVSII-1(G>A) and the cryptic splice site mutation IVSI-110 (G>A) (Table 1-2). Both mutations result in defective RNA processing, and the recovery of the correctly spliced

products can be assessed following gene activation using CRISPR/Cas9-mediated protein targeting with transcription coactivators (see further discussion in Section 1.5). To date, footprint-free iPSC derivation from beta thalassaemia patients with the homozygous mutation IVSII-1(G>A) has not been reported (Table 1-1).

1.3.2 Overview of nuclease-mediated gene targeting in human iPSCs

Pluripotent stem cells offer an ideal platform as they possess indefinite self-renewal capacity coupled with pluripotency, thus enabling the derivation of all cell types through directed differentiation. Footprint-free iPSCs may not only be useful in addressing the quantitative obstacle of expanding peripheral blood-derived HSCs [50], but may also provide a safer approach (in contrast to the viral gene therapy) for correcting the disease-causing mutation. However, gene targeting by homologous recombination (HR) in hiPSCs has been reported to be technically difficult, with an inherently low efficiency [51]. Recent advances in genetic engineering, namely the use of site-specific nucleases that stimulate the endogenous DNA repair mechanisms, can significantly improve the efficiency of gene correction, and are thus increasingly used to overcome this bottleneck [52, 53]. The efficiency and applicability of the system to a wide range of cell types has already facilitated the generation of novel cellular and animal models of disease [8]. However, despite the rapid advances in this field, data on the fidelity of targeted modifications (off-target assessment) at endogenous genomic loci remain highly variable [54, 55]. This is likely related to the fact that the consequences of nuclease activity are modulated by a complex interplay between DNA target accessibility, the kinetics and the nature of nuclease target recognition cell-type specific DNA repair mechanisms and the cell cycle.

1.3.2.1 *Site-specific endonuclease platforms*

Several different classes of endonucleases with programmable specificity have been used for genome editing [8, 52, 53]. Generally, they can be grouped into four main classes: meganucleases, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and clustered regularly interspersed short palindromic repeat (CRISPR/Cas9)-based nuclease systems. Meganucleases are formed from one protein domain that contains both recognition and cleavage functions but possesses limited targeting potential. In contrast, the ZFNs, TALENs, and CRISPR/Cas9 systems are composed of two domains; a programmable domain that directs the nuclease to bind to the intended DNA locus and a nuclease domain that cleaves chromosomal DNA. The programmable domain in the meganucleases, ZFNs and TALENs is a protein, whereas the CRISPR/Cas9 systems rely on RNA target specificity, hence the name RNA-guided nuclease (RGN) [56, 57]. A comparison between the major characteristics of some commonly used nuclease

platforms in pluripotent stem cells is shown in Table1-3. Although both RNA and protein types of DNA-recognition domains can be engineered to bind to a specific target sequences, RNA-based targeting offers a versatile system, with easier design, is amenable to multiplexing, and in principle allows the targeting of any given genomic locus, even if the locus is methylated [58].

Table 1-1: Method of derivation and source of iPSC lines generated with beta thalassaemia mutations

Cell type	Vector	Gene correction strategy	Allele1	Allele2	Reference
Fibroblast	Retroviral	Not done	CD 41-42 (-TCTT)	CD41-42 (-TCTT)	[28]
		Homologous recombination			[26]
	Episome	TALEN Cas9WT			[37, 38]
Fibroblasts	Sendai virus	Cas9WT	CD41-42 (-TCTT)	-28 (A/G)	[35]
Fibroblasts	Lentivirus (STEMCCA), excision not mentioned	Cas9WT	CD17 (A>T)	CD17 (A>T)	[36]
Fibroblasts or bone marrow MSCs	Lentivirus	Gene addition	CD39 (C>T)	CD39 (C>T)	[27]
			IVSI-110 (G>A)	CD39 (C>T)	
			IVSI-1 (G>A)	CD39 (C>T)	
			IVSI-5 (G>C)	CD39 (C>T)	
Amniotic fluid cells	Episome	TALEN	IVSII-654 (C>T)	IVSII-654 (C>T)	[38]

Table 1-2: List of beta thalassaemia mutations (in iPSC lines) and the mechanism by which they affect β-globin chain synthesis CRISPR/Cas9 platforms

Mutation	Stage of gene expression affected	Mutation category
-28 (A/G)	Transcriptional mutation	Promoter element
IVSI-110 (G>A)	RNA processing	Cryptic splice site
IVSI-1 (G>A)	RNA processing	Splice junction
IVSII-1(G>A)	RNA processing	Splice junction
IVSI-5 (G>C)	RNA processing	Consensus splice site
IVSII-654 (C>T)	RNA processing	Cryptic splice site
CD 41-42 (-TCTT)	RNA translation	Frameshift
CD17 (A>T)	RNA translation	Nonsense
CD39 (C>T)	RNA translation	Nonsense

1.3.2.1.1 Cas9 nuclease structure and mechanism

The CRISPR/Cas9 nuclease systems are a new class of versatile, programmable nucleases that utilise an RNA-mediated editing process to enhance genome engineering [59]. The *Streptococcus pyogenes* (Sp) type II A CRISPR/Cas9 system is one of the most commonly used nucleases in genome engineering in mammalian cells. It originally functioned as part of the bacterial adaptive

immune system, providing acquired resistance against invading foreign DNA sequences. The system was next adapted to recognise and target specific DNA sequences (protospacers) in mammalian cells [60]. The recognition process relies on using a complementary sequence (crRNA) that requires the presence of a noncoding trans-activating crRNA (tracrRNA) sequence for efficient Cas9 nuclease activity [59]. The tracrRNA:crRNA structure can be used in the form of either a dual [61] or a chimeric single guide RNA (sgRNA) [58, 62]. The latter can include a 20-nucleotide crRNA target sequence within a ~100-nucleotide RNA sequence of the fusion transcript where the tracrRNA sequence acts as a scaffold. The RNA hybridises to a predetermined complementary DNA sequence and directs the Cas9 endonuclease to bind and cleave the matched DNA site provided that the sequence is followed by a PAM [60, 62-64]. The PAM motif sequence is predominantly composed of NGG (N is a degenerate base), or to a lesser extent NAG, NGA, NCG, NGC, NGT, NTG, and NAA. The alternative non-canonical motif sequences have been identified using *in vitro* cleavage assays, a reporter system, or through an unbiased sequencing method that captures double-stranded oligodeoxynucleotides into DSBs (GUIDE-Seq) [58, 65, 66].

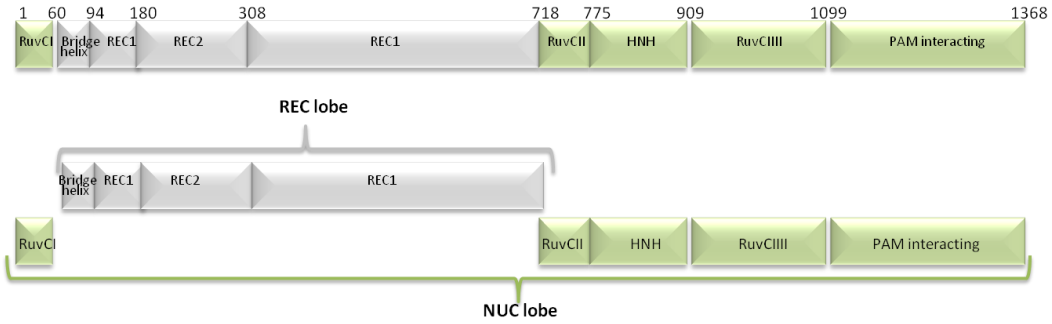

Figure 1-2: Schematic of Cas9 protein domains.

The schematic show the bilobe (NUC and REC) and the six domains of Cas9: Rec I, Rec II, Bridge helix, RuvC, HNH, and PAM interacting. Numbers indicate amino acid residues. Figure adapted from [67].

Structural analysis of the Cas9 protein revealed the presence of a cleaving (NUC) and recognition (REC) 'lobe' (Figure 1-2). The latter interacts with sgRNA and the DNA that resides in a positive groove between both lobes in non-specific manner [68]. The interaction was postulated to stabilise the non-target strand of the targeted DNA site (Figure 1-3.E). More recently, protein engineering has been used to successfully modify amino acids at the interface between Cas9 and the DNA, which enhanced cleavage specificity [69, 70].

The modification enhanced targeting specificity by destabilising the non-specific interaction between the Cas9 protein and the DNA strand, rendering the binding more dependent on the base

pairing stringency between RNA and the DNA sequence (Figure 1-3.E). This approach led to unfavourable binding at off-target sites, due to the presence of sequence mismatches [69]. Recent evidence suggests that the Cas9 system varies from other types of nucleases in its binding kinetics, in that the protein remains bound to the targeted locus after cleavage [71, 72]. The number of cleaved DNA products induced by Cas9 reaches a maximum in one minute, with no further increase with time. This same trend was observed over a range of enzyme concentration (2.5 nM–100 nM), indicating that Cas9 is a single turnover enzyme that does not follow a Michaelis-Menten kinetic model [72].

Table 1-3: Comparison of the major characteristics of some commonly used nuclease platforms in genome engineering

	ZFN	TALEN	Sp CRISPR-Cas9
DNA recognition domain	Protein	Protein	RNA+PAM motif
	Zinc fingers DNA binding motifs	Transcription activator-like effectors	
	Each motif recognises three bases	Two amino acids in the repeat variable di-residue in the DNA binding repeat domain dictate the base preference	Watson-Crick base pairing between DNA and crRNA
	Tuneable by changing the number of DNA binding modules		Truncation of sgRNA to 17 bp or appending 5′-GG may improve specificity
Recognition sequence	9–18 bp/monomer	14–20 bp/monomer	17–20 bp
Nuclease domain	Type IIS restriction enzyme FokI	Type IIS restriction enzyme FokI	Cas9
Methods used in engineering the DNA binding site	Protein engineering	Complex molecular cloning	Standard molecular cloning/cloning-free
Type of DNA cleavage produced	Cohesive ends: 5′ overhangs at cleavage sites		Blunt ends (common) 1 or 2 overhangs *in vitro*
Nickase variant	√	√	√
Sensitive to methylation	Possible	√	
Repurposed for gene regulation	√	√	√
Used in clinical trial	√		

Cas9 nuclease induces mostly blunt DSBs through the activity of two domains that independently nick the complementary strand (the NHN domain) and the non-complementary strand (the RuvC domain) [59]. Both domains are in the NUC lobe of the Cas9 protein. Interestingly, small overhangs were also reported during *in vitro* studies, which were attributed to exonuclease activity at the non-complementary strand [59]. DSBs stimulate the cellular repair mechanisms (see Section 1.4 for details) through either NHEJ or homology-directed repair (HDR) in the presence of a donor template.

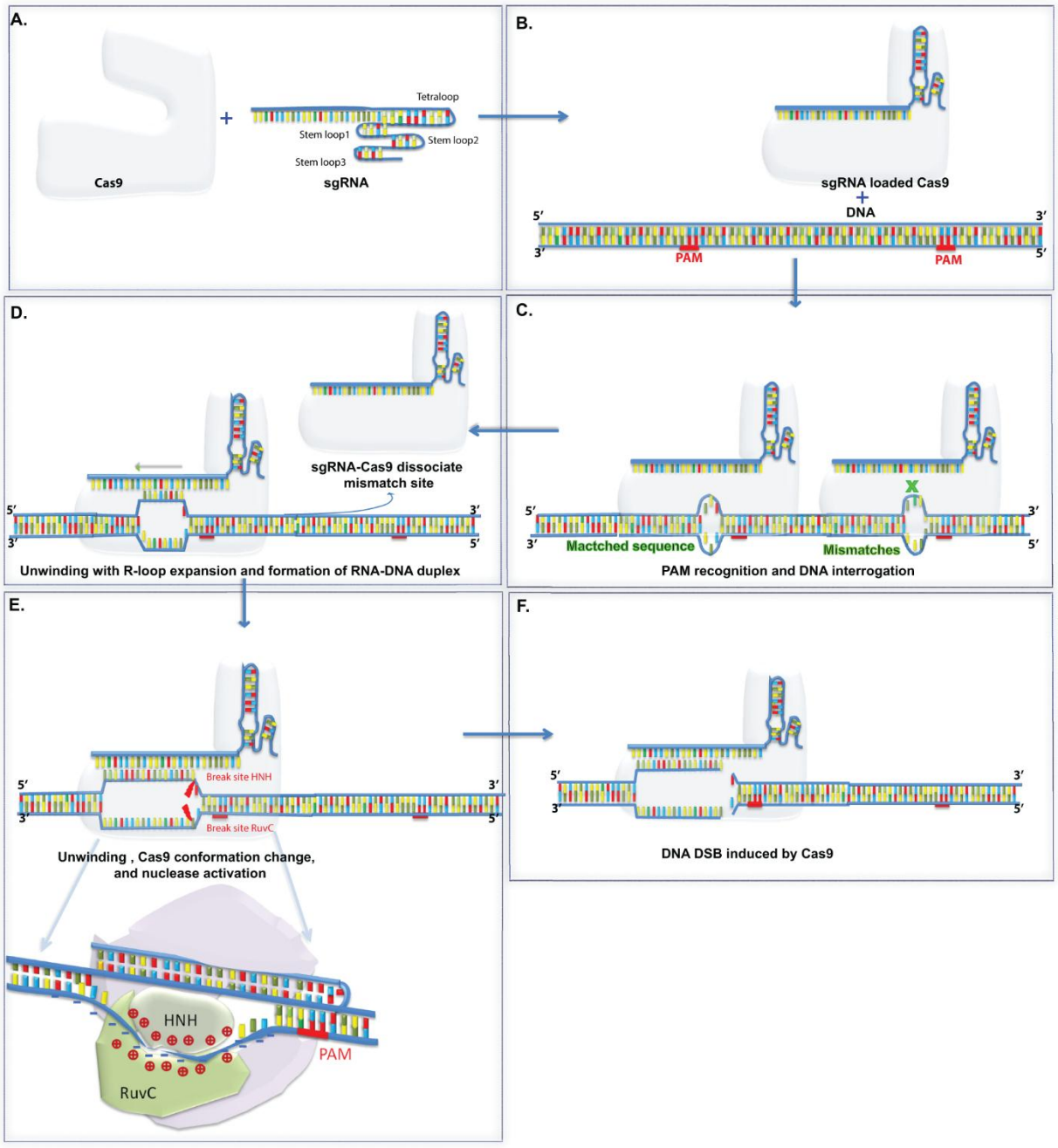

Figure 1-3: Schematic illustrating the process of inducing DSB at targeted DNA locus. A) Association of hybrid sgRNA to the Cas9-protein leading to conformational changes in the latter. B) The REC lobe of the formed ribonucleoprotein complex (sgRNA-Cas9) searches the DNA sequence and recognises a PAM motif, resulting in limited unwinding of the DNA duplex upstream of the PAM, forming a small bubble. C) A RNA-DNA heteroduplex is initiated; however, mismatches destabilise the structure leading to sgRNA-Cas9 dissociation (D). E) Upon base pair matching the loop is extended and the structure is stabilised by nucleotide base pairing and the non-specific interaction between the negatively charged non-target strand and the positively charged groove located between the HNH and RuvC domains. F) A proper base pair match occurs and loop extension induces a conformation change that activates the nuclease activity and introduces a DSB. The complex does not immediately dissociate from the binding site following cleavage. Figure adapted from [69, 72, 73]. PAM sequences are highlighted by red lines.

NHEJ is mostly an error-prone mechanism that results in mutagenic end joining, whereas HR leads to a defined and precise modification of a genomic DNA sequence dictated by the sequence in the donor vector. Consequently both types of DNA repair pathways have been exploited to facilitate the generation of gene knock-outs, gene or transgene knock-ins and point mutation.

1.3.2.1.2 Off-target activity and strategies to enhance specificity

Tolerance to several mismatches between guide RNA and DNA sequences raised concerns about undesired off-target cleaving that led to mutation as a result of repair through the NHEJ pathway, with possible undesired consequenses [58, 74, 75]. Remarkably, off-target activity was not limited to mismatches (base substitution). In fact, a DNA-RNA bulge formed as a result of the presence of insertions or deletions in the sequence of the off-target locus was also tolerated by Cas9 [76]. This non-specific activity potentially limits the utility of the Cas9 system for both basic research and therapeutic clinical translation. Although the determinants that govern the specificity of Cas9 are not fully characterised, various recommendations and strategies have been suggested to address this limitation. Details of these recommendations that include guide sequence properties and approaches to improve specificity will follow in the next section.

Guide sequence properties that can play a role in modifying Cas9 specificity include:

a) Unique target site selection: Li *et al.*[77] showed that enhanced specificity can be obtained by designing a sgRNA to target a DNA site with unique sequence. This study provided an analysis tool (software) that precluded sequences with mismatches by mapping all possible substrings of length sequence with less than 16 bases to the genome. The increased depth displayed by the mapped sequence reflects a possible unique target site (analysis of the *HBB* gene with the software provided is shown in Figure 1-4). Evidently, this strategy somewhat limits the selection of gRNA in gene correction studies as the placement of the DNA break relative to the intended modification site is critical for the efficiency of gene correction [78].

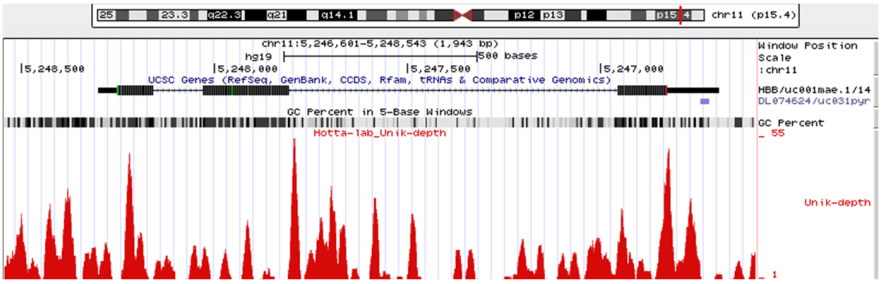

Figure 1-4: Unique k-mer to identify a unique sequence in the HBB gene. (https://apps.cira.kyoto-u.ac.jp/igeats) [77].

b) Selection of sequences with predicted off-target sites (DNA sequences with few mismatches) in a closed chromatin structure (heterochromatin with H3K27me3 mark) [76]: Although early studies indicated that the chromatin state does not inhibit the accessibility of Cas9 [58, 78, 79], recent evidence suggests that the chromatin environment may impact on Cas9 activity, as indicated by the fact that DNA in nucleosome structures was reported to be less prone to nuclease recognition and binding [80]. However, it should be noted that the dynamic nature of nucleosome occupancy and turnover during replication can render these DNA sites available for cleavage [80]. Nevertheless, others report that: (1) the search efficiency of Cas9 within heterochromatin regions displays slower kinetics in comparison to that in euchromatin regions [81]; (2) the genome-wide binding profile of Cas9, as revealed by Chip-seq analysis, shows that Cas9 displays a preference for open chromatin areas [82]; (3) discrepancies have been observed in the number of off-target sites identified in the cellular genomic context as compared to that in the *in vitro* context (no chromatin, Digenome-Seq) [83]. The number of off-target sites predicted using the *in vitro* approach were significantly higher than those captured in a cellular genomic context [54, 83, 84]. Yet, it is not clear if this is solely attributed to the differential chromatin state. Overall, the evidence supports a role for chromatin accessibility in Cas9 activity. Indeed, one of the recent bioinformatic tools for off-target prediction integrates information on the chromatin state into the analysis [85]. In the context of this thesis, it is important to note that pluripotent stem cells adopt an open chromatin structure that differs from that of somatic cells [86], implying that iPSCs may have a higher propensity for off-target activity. Surprisingly, therefore, off-target Cas9 activity in iPSCs was reported to be low [87] or non-existent [88, 89]. However, it should be noted that Tan *et al.* [87] limited their screen to an analysis of ~90 off-target sites for 12 gRNAs, whereas Smith *et al.* [89] and Veres *et al.*[88] used whole genome sequencing in their study, for only a limited number of sgRNAs, in iPSCs. This implies that the interplay of various factors determines Cas9 activity, including cell-type specific parameters that influence off-target activity and can potentially explain these apparently discordant findings, as discussed in Section 1.4.

c) Modulation of the amount of sgRNA-Cas9 transfected: Off-target activity may be able to be reduced by limiting the amount of sgRNA-Cas9. One way to achieve this is by introducing a uracil-rich seed sequence (proximal to PAM) in order to decrease the abundance of sgRNA by affecting RNA transcription and thus improve specificity [82]. Similarly, titrating down the amount of sgRNA-Cas9 decreased non-specific cleavage in one study [58], but was not effective in changing off-target activity in another [74]. This latter observation was supported by a study by Frock *et al.* [90]who did not observe a change in either the qualitative (number of sites) or

quantitative (activity levels) off-target activity over a 10-fold range in the amount of sgRNA-Cas9 (for a single locus). Evidently, the role that decreasing the amounts of Cas9 and sgRNA has in limiting off-target activity remains controversial possibly depending on guide sequence and locus context properties.

d) Altering the stability of Cas9 binding: The GC content of the guide was also suggested to play a role in both off-target [74, 76], and on-target, activity [91]. A higher GC percentage is generally associated with higher activity. This was attributed to an increase in the stability of Cas9 binding associated with guide RNAs rich in GC nucleotides. Indeed, 28 per cent of the variation seen in Cas9 efficiency in inducing mutagenesis at the targeted locus was attributed by the melting temperature of the sequence (gRNA) [78]. The stability of binding may explain some of the differences observed between studies investigating the effect of titrating the amount of sgRNA-Cas9 on off-target activity; in other words, guides with high stability can produce higher off-target activity with little effect related to the available amount.

e) Manipulation of the guide structure: Altering the guide structure by either decreasing the length of the guide (truncated guides) [92], or appending GG bases to the 5' end of the guide [61], may promote specificity. The concept of decreasing the guide length was based on a thermodynamic model which predicts that shorter guides would be more prone to destabilization by nucleotide mismatches and thus increase the fidelity of the system [92].

Alternatively, enhancing specificity can be achieved by using different forms or variants of Cas9:

a) Transient expression: Transfection of Cas9 ribonucleoprotein was used instead of the plasmid form in order to avoid prolonged expression of Cas9 [93, 94]. This was based on the observation that the ribonucleoprotein was cleared from the cells within 24 hrs whereas plasmid-delivered Cas9 displayed continuous expression for ≈72 hrs.

b) Engineer mutations in the Cas9 nuclease to generate variants with properties that limit off-target activity: One strategy explored a cooperative model for DSB generation. (1) Converting the nuclease into a nickase variant (Cas9n) by introducing either D10A, H840A (or N863A) amino acid substitutions into the active site of the RuvCI or the HNH nuclease domains, respectively, effectively inactivating one catalytic domain [68, 95]. Inactivation of DNA cleavage at one strand thus allowed the introduction of an SSB, instead of a DSBs, at the targeted DNA site. Commonly, the Cas9D10A variant is used in conjunction with two guide RNAs to generate a composite DNA DSB on opposite DNA strands within a defined offset range at the targeted locus [96]. This so-called 'paired nickase' strategy aims to improve the on-target cutting efficiency of the nickase by inducing a composite DSB at the targeted locus

(double nickase), while SSBs are induced at off-target sites by a single nickase. Mutagenesis at such off-target sites would be minimised because nicks are thought to be repaired through high fidelity repair mechanisms [96, 97]. (2) An alternative strategy involve the conversion of Cas9 into a deficient nuclease protein (dCas9), which was accomplished by inactivating both cleaving domains (HNH and RuvC). The inactive Cas9 protein was next fused to non-specific wild-type FokI endonuclease. The specificity of this system is dictated by the synergistic interaction of two monomers of the fused protein (dCas9-FokI), where sgRNA directs the binding of dCas9, and the FokI endonuclease enables cleavage of the site upon dimerisation [98, 99]. The system is analogous to the ZFN or TALEN models. Of note, combining FokI-dCas9 nucleases with truncated guides has also been reported to further increase the specificity of DNA cleavage [100]. A second strategy aims to weaken the non-specific interactions between Cas9 and either the on-target strand (eSpCas9) [69] or the DNA phosphate backbone (SpCas9-HF1) [70] in order to destabilise the binding at non-specific targets.

Although the majority of the above-mentioned approaches were reported to enhance the specificity of DNA cleavage, the effectiveness of these approaches varied between different DNA loci and cell type. For example, while the truncated guide was reported to increase specificity of sgRNA-Cas9 mediated DSB, as assessed by a genome-wide approach (Guide-seq) [65], another study reported persistent off-target activity with truncated guide RNAs, as well as the creation of new off-target sites compared to the full-length 20-mer counterpart [69]. Similarly, while the paired nickase strategy was proven to enhance the specificity of DNA cleavage in mouse embryos, and no mutagenesis associated with the monomeric nickase (Cas9D10A) was reported [101], other studies reported either low level of on-target activity associated with monomeric Cas9n in immortalised or transformed human cells lines [61, 98], or a tolerance of the RNA-DNA bulges at off-target sites [76]. It is clear, however, that nickase (or Cas9DN) certainly enhanced specificity when compared to Cas9WT [90, 96, 102]. As discussed previously, it remains difficult to perform a meaningful comparison between these different strategies as discrepancies may reflect differences in variables such as guide sequence properties [76, 91], differing states of local chromatin [80], the format of the transfected sgRNA-Cas9 (plasmid, mRNA, or ribonucleoprotein) [93, 94], the amount of both guide and Cas9 used, the sensitivity of the method used to detect mutagenesis (Surveyor, T7endonuclease, Digenome-seq, Guide-seq ect) and cell type. In this thesis, the double nickase approach was adopted and optimised for gene correction of the splice junction mutation in beta thalassaemia-iPSCs.

1.4 *Repair mechanisms and pluripotent stem cells*

Cell type, cell cycle phase, the type of DNA damage, and chromatin state all influence a cell's response to DNA damage [103, 104]. Pluripotent stem cells have certain unique characteristics that differ from other cell types; for example, different cell cycle phase lengths that are associated with the pluripotent state of the cell [105, 106], lack of some cell cycle checkpoints in response to certain types of DNA damage-inducing agents [107] and an open chromatin structure [86]. These properties may result in different repair kinetics for similar target sites in different cell lines. In the following sections, the major aspects of the DNA repair mechanisms and their relevance in resolving DNA breaks induced by nucleases in iPSCs will be outlined.

1.4.1 Overview of DNA repair mechanisms

The integrity of the genome is preserved by the concerted action of several cellular repair mechanisms commonly referred to as the 'DNA damage response' (DDR). The DDR includes: nucleotide excision repair (NER), base excision repair (BER), mismatch repair (MMR) and recombinational repair [108]. SSB breaks are mostly repaired through BER [97] while both SSB and DSB can be repaired via the recombinational repair pathways in the G2 phase of cell cycle [97, 109]. The recombinational repair pathways include four pathways that fall into two different classes depending on the presence, or absence, of resection processing at the DNA break site (see Figure 1-5). The resection-dependent pathways include alternative end joining (A-NHEJ, mutagenic), single-strand annealing (SSA/micro-homology-mediated end joining (MMEJ), mutagenic) and the homologous recombination pathway (HR, high fidelity). The resection independent pathway involves classical non-homologous end joining (C-NHEJ) and is mutagenic, but can also be error-free [108, 109].

Following detection of DSBs, the DNA site is marked by a molecular signal with subsequent recruitment of the repair machinery. In C-NHEJ, DNA resection is prevented by activating accessory factors that protect the break site [108]. Alternatively, in the resection-dependent pathways, nucleolytic degradation at the 5' end of the DNA at the break site is stimulated [108, 109]. As a result, ssDNA with 3' overhangs is generated through the action of double-stranded DNA nuclease, exonuclease and helicase (see Figure 1-5 for further details on the mechanism). The resection is limited to 5-25 bp for A-NHEJ but is broader in HR and SSA [110]. The HR and BER pathways are considered high fidelity pathways that restore the integrity of the DNA sequence, whereas A-NHEJ, SSA, and C-NHEJ are generally considered to be error-prone pathways [108, 109, 111]. Achieving high fidelity repair in nuclease-mediated gene targeting of PSCs relies on attenuating the mutagenic pathways and shifting the balance to high fidelity repair pathways (i.e.

HR and BER). Indeed, molecules that either inhibit NHEJ, or enhance HDR, were reported to enable more efficient and precise gene editing [112, 113]. Although many aspects of the regulation, mechanistic details, and repair pathway choices remain to be elucidated, several contributing factors that may play a role in coordinating the choice of repair path have been defined [109, 114].

The cell cycle has been reported to be one of the factors determining the repair path followed (resection-dependent/independent) (Figure 1-5). Recent data indicates that HR is actively repressed during G1 phase [115]. This effect was mediated by preventing the assembly and localization of a protein complex (BRCA1–PALB2–BRCA2) required to promote HR at the DSBs during G1. HR predominates in the mid S/G2 cycle phases, where a homologous template is available [109]. On the other hand, data on the BER pathway are inconclusive. Caldecott [97] indicated that BER was active at all cell cycle phases, while Dianov and Hubscher [111] reported that it is only active in the G1 phase of the cell cycle. More recently, Mjelle *et al.* [116] analysed six transcriptome data sets from synchronised human cells, revealing that the expression of genes related to short patch BER are present throughout the cell cycle, while those associated with long patch BER were limited to the S phase. NHEJ was reported to be active throughout the cell cycle, with dominance in the G0/G1 phase, and is independent of homologous sequence [114]. Both cooperation and competition between HR and NHEJ may occur at cell cycle phases where both pathways are active; (i.e. G2-phase), in a manner that maximises repair outcomes and genomic integrity (an outline of the major repair pathways and activities in a cell cycle context are shown in Figure 1-5) [114].

Different types of DNA breaks favour different repair mechanisms. DSBs and SSBs both lead to the stimulation of DNA repair. DSBs are mostly repaired through the recombinational pathway (NHEJ or HR) [109], while SSBs are thought to be predominantly repaired by dedicated pathways such as BER [97, 111]. However, SSBs encountered during DNA replication leads to the generation of DSBs due to collapse of the replication fork, and will thus shuttle the repair to DSB resection-dependent pathways [97]. In this respect it is intriguing to note that HR at nicks introduced by nuclease variants does not seem to be triggered by the formation of DSB intermediates [117, 118]. In fact, various studies have indicated that nickase stimulates HR, yet the details of this mechanism remain to be elucidated [118, 119].

Programmable nucleases, such as Cas9, induce either a DSB or a SSB depending on the protein variant used. The DNA termini (3' or 5') generated by Cas9 are not damaged and therefore may facilitate the restoration of the sequence by religation (i.e. C-NHEJ or BER) without end resection [97, 120]. Translocation through precise end joining was demonstrated for blunt end DSBs induced by Cas9WT [121]. Moreover, a recent study demonstrated a high percentage of precise end joining between two blunt end DSBs that generated genomic microdeletions in different cell types [120],

lending further support to the idea that blunt ends generated by Cas9WT can be repaired through error-free NHEJ. It should be noted, however, that in the study by Geisinger *et al.* [120] deletion of the intervening sequence would have prevented the iterative cycle of cleavage and repair induced by the nuclease [114]. Upon further examination of the data in that report, the range of precise ligation was higher in H9 pluripotent stem cells (>75 per cent, except for one guide RNA) than in HEK 293 cells (which varied between 44 per cent and 90 per cent), despite a similar repair pattern in both cell types.

It can be concluded that DNA breaks induced by CRISPR/Cas9 can be repaired through either error-free pathways that include BER, and HR, or error-prone pathways, but that the choice of pathway and its activity is dependent upon the locus, cell cycle, and the type of DNA break. Furthermore, As has already been discussed in the previous paragraph, NHEJ can result in an error-free DNA repair [120].

1.4.2 DNA repair in pluripotent stem cells

PSCs possess several unique cell and DNA repair characteristics. Mouse ESCs (mESCs) have been reported to favour HR [122] and exhibit a low level of NHEJ activity [123]. The majority of pluripotent stem cells exist in S phase, where resection-dependent mechanisms are active as opposed to C-NHEJ [106]. This may explain the observed decrease in mutation frequency in mESCs compared to their somatic counterparts [124]. A low expression level of some of the NHEJ-related proteins was also found in mESCs [123].

Although species differences in PSC are well documented [121, 125], HR was suggested to predominate in hESCs, similar to mESCs [126]. When compared to other human cell lines, differentiated cells derived from hPSCs, hPSCs display an enhanced DNA repair capacity following exposure to various DNA-damaging agents. They also exhibit higher expression levels of genes and proteins involved in the repair mechanisms [127-129], and a lower level of NHEJ-mediated repair of DSBs [130]. Using human fibroblasts, a cell type in which C-NHEJ is the primary choice of repair mechanism [131], Shibata *et al.* [131] showed that the speed of repair influences the repair path chosen and that this depends on both the chromatin and DNA damage complexity, so that a slower repair may shuttle the repair through the resection-dependent pathway. Consequently, it is reasonable to infer that the relaxed chromatin environment in PSCs may predispose the cell to different kinetics for its DNA repair pathways [132, 133].

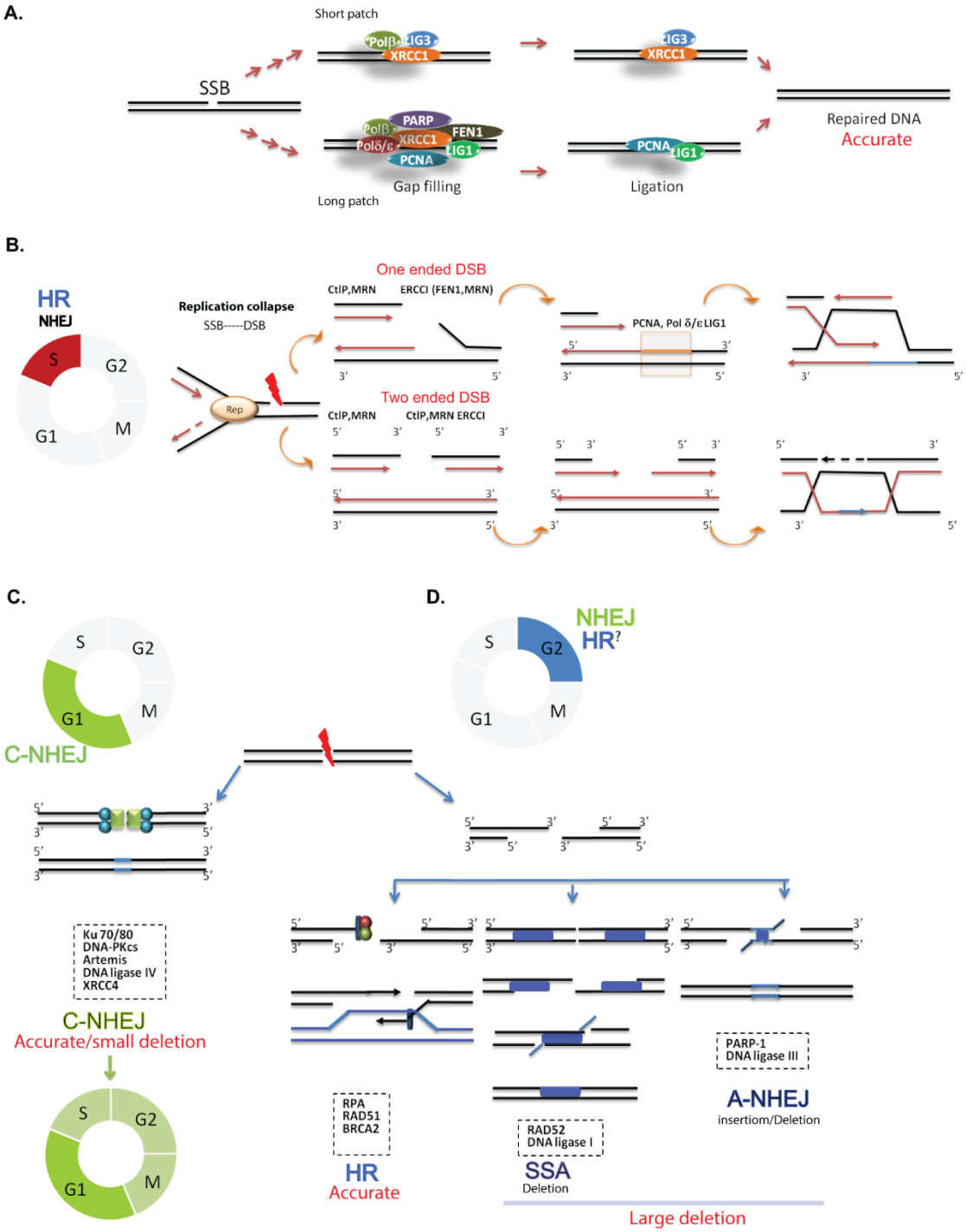

Figure 1-5: Schematic for the major repair pathways for SSB and DSB. (A) Specialised proteins are recruited at SSBs (depending on the type of DNA damage) and end processing of DNA termini ensues, followed by gap filling through either short patch or long patch repair and ligation. (B) SSBs that escape repair in G1 and progress to S phase lead to replication fork collapse and the possible formation of either one-ended or two-ended DSB. DSB are mostly repaired through accurate HR or through NHEJ. (C) During G1 phase DSB are mostly repaired through NHEJ, resulting in either precise or small deletions. NHEJ is active throughout the cell cycle (D); however, HDR and C-NHEJ compete in G2 phase, and the repair kinetics are mostly determined by the type of DNA damage and chromatin complexity. Upon activation of resection HR, SSA and A-NHEJ compete for DSB repair. Figure adapted from [97, 109, 134].

A separate mechanism that maintains genome integrity and prevents the accumulation of DNA damage in PSC populations is apoptosis. In response to DNA lesions, PSCs display an increased propensity to activate apoptosis compared to that of somatic cells, and this propensity is lost following their differentiation [135]. Activation of apoptosis following DNA damage, as opposed to activation of critical cell cycle phase checkpoints during replicative stress to enhance genomic integrity, constitutes a protective mechanism that prevents the accumulation of cells with genomic aberrations in some cell types (such as HSCs and PSCs) [134]. Indeed, in contrast to most cell types, both hESCs and hiPSCs undergo apoptosis and fail to activate the S phase checkpoint (G1/S, CHK1) following exposure to replication inhibitors and gamma radiation [107, 136]. At odds with these data, a functioning checkpoint was reported upon exposing pluripotent cells to DNA-damaging ultraviolet radiation in the C range (UVC) light [137]. Overall, the preponderance of evidence indicates that the DNA repair response to the Cas9 system in PSCs is likely to lead to very different genomic integrity outcomes in iPSCs compared to other human cell lines.

1.4.2.1 *Evidence for DNA repair at off-target sites in iPSCs*

Data regarding the resolution of DNA break induced by Cas9 at off-target sites in PSC vary greatly between studies. Whole genome sequencing revealed that Cas9WT showed a highly specific cleavage activity in pluripotent cells. Few off-target indels were identified, and none of them showed sequence homology with the targeted site or were recurrent between the screened clones [88, 89]. However, these studies were carried out on a limited number of clones using a small number of targeted sites. Further studies confirmed the absence of nonspecific activity by screening for cleavage activity at the most likely predicted off-target sites based on sequence homology to the guide RNA sequence [35, 36, 77]. Moreover, upon screening 624 and 717 potential off-target sites for 12 sgRNA in 293FT cells and iPSCs, respectively, a total of three sites were identified as being associated with two sgRNAs only [87]. On the other hand, Xu *et al.* [138] showed clear off-target activity at seven out of 10 sites after targeting Intron 2 of the *HBB* gene, and reported a comparable level of off-target activity in 293FT cells and iPSCs. The discrepancy between these studies may be related to the sequence of the sgRNA used and the local environment of the targeted locus.

In conclusion, although the overall evidence points to possible high-fidelity repair of DSBs in PSCs, DSBs have the potential to be repaired via mutagenic repair pathways, with potential consequences resulting from off-target activity. The prediction of possible off-target activity is challenging, as it can be locus- and sequence (sgRNA) -dependent. Mutagenic end joining is rarely associated with SSBs or nicks. Evidence indicates that SSBs can predominantly stimulate HDR, albeit with markedly reduced efficiency compared to the DSBs introduced by the wild-type Cas9

nuclease [118, 119]. This strategy would be especially relevant for genes such as *HBB*, which has been reported to have a considerable propensity for off-target activity [139]. The number of studies that have investigated the validity of different strategies for enhanced specificity of gene correction in PSCs, and that have comprehensively analysed on-/off-target activity, is limited. This thesis attempts to address some of this deficiency.

1.4.3 Gene editing in beta-haemoglobinopathies

A number of studies have successfully used programmable engineered nucleases such as ZFNs, TALENs and RGNs for targeted repair of the disease-causing mutation in iPSCs derived from patients with monogenic disorders affecting the β-globin gene, including sickle cell anaemia [140-143] and thalassaemia [26, 35, 36, 38, 138, 144]. In contrast to ZFNs and TALENs, the Cas9 system offers the advantage of a higher targeting efficiency using a cost-effective, flexible and versatile system [8]. The genotoxicity analyses in studies that used Cas9-based systems for gene correction of the beta thalassaemia mutation are mostly limited to a handful of computationally-predicted potential off-target sites that bear high sequence homology to the on-target site. A high rate of off-target activity (7 sites with mutagenesis out of 10 screened sites) was reported in one of those studies [138]. Recent evidence suggests that bioinformatic methods may not accurately predict off-target sites, given the complex milieu of the factors determining functional off-target sequences [65]. Since precise genome editing is required to realise either the therapeutic or the disease-modelling potential of iPSCs, accurate quantification of off-target, as well as on-target mutagenesis by Cas9 remains an important issue in the field.

As discussed above, any endonuclease that generates DSBs is associated with potential off-target activity that can compromise the specificity of the system. To reduce the potential for adverse mutagenic off-target activity by Cas9, a nickase variant (Cas9n) of Cas9 was developed. Because the nickase activity is reported to have low on-target efficiency, a double nicking strategy was devised that can improve this by inducing composite nicks [96]. Testing the applicability of the paired nickase strategy, which is predicted to have higher specificity compared to the native CRISPR-Cas9, for targeted gene correction of the *HBB* mutation is one of the main objectives of this thesis. Because the activity of the endonuclease (nickase or double nickase) and its induced HDR depends on different variables that include the targeted locus, gRNA sequence, and type of cells used [117, 145-147], it was expected that this strategy would require extensive optimisation. Furthermore, the effectiveness and fidelity of this system can also be influenced by the choice of nick sites relative to the mutation site or the insertion site [96, 145]. For instance, the design strategy can influence the occurrence of insertions/deletions at the targeted site, since the

continuous activity of the nuclease is likely to reduce the corrective editing rate [145]. Although the study [145] has suggested that physical separation of the guide pair by an insert can attenuate undesired on-target mutation following gene correction, this strategy has not been evaluated in the context of correcting a point mutation. Furthermore, testing of an alternative strategy that introduces a silent mutation designed to prevent repetitive cleavage cycles at the corrected allele was also carried out.

1.5 *CRISPR/Cas9-mediated HBB activation*

Following gene correction in disease iPSCs, validation of the restoration of gene function is essential. However, if the gene is silent in iPSCs, directed differentiation to specialised cells that express the gene of interest is required. In the case of the haemoglobinopathies, this presents a challenge since little expression of either the β-globin RNA transcript or protein has been reported following differentiation of pluripotent stem cells into the erythroid lineage. Furthermore, PSCs display defective haemoglobin switching, with incomplete erythrocyte maturation and dominant expression of the embryonic-foetal hemoglobin phenotype [148-152]. A low levels of β-globin switching have recently been reported [143, 153]. This was partially attributed to the use of blood-derived iPSCs that may have an improved differentiation potential towards the haematopoietic lineage because of epigenetic memory. Although, a recent study has shown otherwise [154]. To overcome this obstacle, Trakarnsanga *et al.* [155] deployed a strategy whereby erythroid cells derived from iPSCs were transduced with two developmental regulators of haemoglobin switching, *KLF1* and *BCL11A*, and this was found to induce an adult level of β-globin. Alternatively, DNA manipulation via null variants of RGN fused with an effector (activators, CRISPRa), could be used to facilitate the modulation of gene output, interrogation of biological processes and dissection of regulatory elements in mammalian cells without cleaving or inducing a mutation in the genome [56, 156, 157]. In this thesis, this approach was explored in beta thalassaemia iPSCs.

1.5.1 CRISPRa-platforms for gene activation

Activation of a gene can be stimulated by various mechanisms. Activation can be mediated by recruitment of transcription factors or by epigenetic factors that facilitate forced DNA looping and/or recruitment of histone modifiers [156, 158]. The potent transcriptional activator domain (TAD) of VP16 (virus protein 16 from *herpes simplex* virus-1) was used in initial Cas9 studies [79, 159]. The activator domain facilitates the assembly of the preinitiation complex at the promoter region of genes and thus upregulates gene expression. Tandem repeats of $VP16_{(n)}$ (TAD) such as VP48 (n=3), VP64 (n=4), and VP160 (n=10) are commonly fused to a DNA binding domain in

order to target the activator to the gene of interest and promote its transcription [156]. VP96 (n=6) and VP192 (n=12) have been used infrequently [160]. ZFN, TALEN and Cas9 systems attached to such tandem VP16 repeats have been successfully used for transcriptional activation of various endogenous genes [79, 161, 162]. The Cas9 system has the additional advantage that it permits multiplexing, achievable by co-expression of multiple gRNAs. Targeted transcriptional activation was successfully achieved by appending p65 [159] or $VP16_{(n)}$ [79, 163, 164] to the inactive Cas9 protein. $VP16_{(n)}$ was used together with multiple sgRNAs to direct the activation complex to the proximal promoter region of various targeted loci. Further, improved efficiency was achieved through recruitment of additional effectors such as the NF-κB, other hetero-effector domains or aptamers (see below) [165, 166]. To date, the Cas9-based system for gene activation can be broadly divided into two main classes depending on the nuclease activity of the Cas9 protein:

The first class involves the use of the nuclease-dead Cas9 (dCas9) which is devoid of endonucleolytic activity. Nuclease null Cas9 can be obtained by inactivating both the catalytic domains (RuvC and HNH domains) of the enzyme [167]. The null variant retains its DNA binding ability which facilitates targeting to specific DNA sites. This modified protein is combined with engineered synthetic transcription factors to form a fusion protein product that is capable of activating the targeted genes. Significant activation of gene was achieved through:

a) A single sgRNA (singleplex) coupled with: (1) A potent engineered chimeric or hetero-activator domain such as dCas9-VP64-p65-Rta, tripartite activator [166], (2) the synergistic activation mediator that comprises VP64 bound to dCas9 and a hetero-effector domain MS2–p65–HSF1 bound to an engineered aptamer-modified sgRNA [165] and the SunTag array, which facilitated the recruitment of multiple copies of VP64 to the targeted site [168].
b) Multiple sgRNAs used in conjunction with VP16(n) (homo-activation domain) system. Multiplexing was used for tiling the proximal promoter region of the targeted gene. Tiling enhanced the low efficiency of gene activation as compared to a single guide [79, 163, 164].

The second class employs a shortened sgRNA in conjunction with an active Cas9 nuclease. The nuclease activity is attenuated by reducing the length of the targeted sequence to 14–15 bp while activation is achieved by appending the effector to either the ‘dead sgRNA’ [169] or to Cas9 nuclease [170]. Dahlman *et al.* [169] achieved 10,000-fold activation for *HBG*1/G2 with Cas9-sgRNA-MS2:P65:HSF1and Kiani *et al.* [170] reported 40 per cent of the expression level observed with dCas9-VPR-sgRNA (14nt) compared to full-length (N20) guide.

Although using hetero-effector domains can mediate efficient gene activation with a single guide, using a $VP16_{(n)}$ system with multiple sgRNA may have the advantage of less off-target activation. In fact, using a single guide in conjunction with the dCas9-$VP16_{(n)}$ system was shown to have a

poor targeted activation efficiency unless it was coupled with tiling across the targeted region using multiple guides [79, 164]. Additionally, a single sgRNA used in conjunction with a potent transcriptional activation may cause off-target activity, as sgRNA has been demonstrated to bind to multiple genomic sites [95]. On the other hand, interrogation of a biological process that requires targeting of multiple genes or using a screening library would clearly benefit from an efficient system that requires only a single guide. More importantly, Cas9-mediated activation is not restricted by the chromatin state of the targeted locus [79]. Therefore, it was considered that this strategy had the potential to enable the assessment of the functionality of the corrected β-globin gene in the absence of erythroid differentiation. Such recruitment of transcription factors to the proximal promoter region of the β-globin gene has, to our knowledge, not been performed before, although targeted demethylation of specific CpGs in the promoter region by a TALE-TET1fusion protein was reported to increase *HBB* gene expression in K562 cells [171]. On the other hand, sgRNA-dCas9-mediated delivery of human acetytransferase p300 to the strong enhancer region of the β-globin region (distal LCR) in HEK293FT with a histone modifier failed to activate *HBB* [172]. Instead, variable levels of upregulation of *HBD, HBG,* and *HBE* were observed. In this study, activation of the *HBB* locus with the CRISPR/Cas9 system in iPSCs was achieved by recruiting transcription factors using the $VP16_{(n)}$ system.

1.6 *Project goals*

The recent explosive advances in genome engineering have provided powerful new tools that not only allow site-specific gene correction, but also, potentially, gene correction *in vivo*. A proof of concept of the latter was recently reported using a mouse model of Duchenne muscular dystrophy in three different studies [173-175]. However, in order to realise their full potential there is a need to explore and address the major challenges attached to the currently available genome editing tools. The establishment of efficient and highly specific gene correction strategies is one important hurdle to overcome. Secondly, the fidelity and off- and on-target effects of such strategies need to be carefully evaluated.

The above discussion highlights evidence that PSCs may have a different on- /off-target profile compared to other cell types due to their unique characteristics. Indeed, review of the major repair pathways that are stimulated in response to DNA breaks indicates that Cas9-induced DNA damage can be repaired through multiple pathways that favour error-free modifications in PSCs. Based on the enhanced capacity of PSCs for repair and their increased sensitivity to apoptosis, it was reasoned that PSCs may have a better potential for either accurate repair or the removal of cells with altered genomes through apoptosis. In this study, it is hypothesised that strategies that convert

DSBs into SSBs, such as a DN strategy, would be more advantageous for precise modification in PSCs. As the efficiency of gene correction is influenced by various factors that include the strategy design, it was further hypothesised that preventing the iterative cycle of Cas9 activity by either separation of the guide pair that targets the locus of interest, or by introducing a synonymous SNP in the PAM motif, would be effective in limiting on-target mutagenesis. A further aim was to assess the applicability of Cas9 variants for stimulating gene expression of the inactive β-globin locus in iPSCs and to explore the potential of this system for modelling the RNA processing defect of the *HBB* gene in beta thalassaemia and determine whether this was corrected following restoration of the wild type allele.

To address these hypotheses, this study aimed to:

1) Generate footprint-free iPSCs from fibroblasts with the beta thalassaemia mutation;
2) Assess the applicability and efficiency of paired nickase strategy to correct the disease-causing mutation;
3) Evaluate the effectiveness and fidelity of two different design strategies for corrective gene targeting in iPSCs;
4) Explore the applicability of the CRISPRa systems for inducing *HBB* transcription activation; and
5) Validate the functionality of the corrected gene in terms of expression of the corrected β-globin gene using CRISPR-based technology.

CHAPTER II:

MATERIALS AND METHODS

2 MATERIALS AND METHODS

2.1 *Establishment of primary cell cultures*

Dermal skin punch biopsies (3 mm^2) were obtained from three patients homozygous for a beta thalassaemia mutation and from one normal control. All primary samples were collected from donors by a trained physician after obtaining informed consent. The β-globin gene mutation was confirmed using the β-globin strip assay MED kit, from Vienna Labs, according to the manufacturer's protocol. Briefly, the DNA was extracted from blood cells and used to amplify a specific region of the β-globin gene in a single multiplex PCR. The biotinylated amplicons were hybridised to an array of probes immobilised on a membrane test strip, allowing the detection of mutant or wild type alleles by a visible enzymatic colour reaction. Two beta thalassaemic patients were siblings with a splice junction mutation (IVSII-1 G>A) that abolishes β-globin chain synthesis ($beta^0$) (Figure 2-1). The third patient had another type of splicing mutation that results in a severe decrease of β-globin protein synthesis ($beta^+$) (IVSI-110 G>A) (Figure 2-1).

Fibroblast cultures were established from patient and control samples using the explant outgrowth method as described by Raya *et al.* [176] with minor modifications. Briefly, a scalpel was used to finely dice the skin punch biopsies. Small pieces were placed in a 60 mm diameter tissue culture plate, covered with a cover slip and cultured in a fibroblast medium consisting of Dulbecco's Modified Eagle Medium (DMEM) (Gibco, Grand Island, NY, http://www.invitrogen.com) supplemented with 10 per cent heat-inactivated foetal bovine serum (HI-FBS), 1x sodium pyruvate, 2 mM Glutamax, 50 U ml^{-1} penicillin and 50 mg ml^{-1} streptomycin. Fibroblast outgrowths were passaged with TrypLE and expanded for two passages before cryopreservation at -80°C in a freezing medium that consisted of 80% FBS, 10 per cent complete DMEM and 10 per cent DMSO.

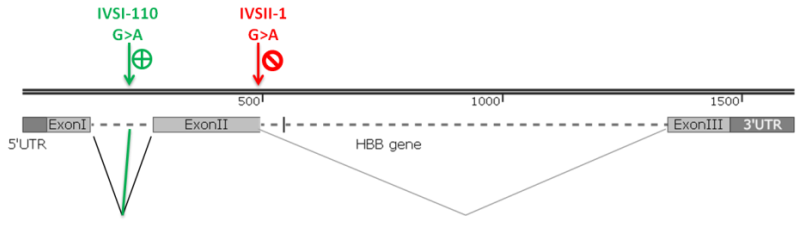

Figure 2-1: Mutations leading to aberrant β-globin mRNA splicing in certain cases of beta thalassaemia. IVSI-110 leads to the formation of a new acceptor splice site whereas IVSII-1 abolishes the splice donor site.

2.2 *Maintenance of cell cultures*

Human embryonic kidney (HEK) 293FT cells (Life Technologies) were maintained in DMEM medium supplemented with 10 per cent FBS, 0.1 mM MEM non-essential amino acids, and 1 mM

sodium pyruvate. Cells were passaged every two days using TrypLE. Pluripotent stem cells were either maintained on mouse embryonic fibroblast (MEF) feeder cells in KSR-hESC culture medium (DMEM Ham's F-12 medium, 20% Knockout Serum Replacement, 1 per cent non-essential amino acids, 2 mM L-glutamine, 0.1 mM beta-mercaptoethanol, and 100 ng/ml basic fibroblast growth factor (all from Gibco, except for bFGF, which was from Invitrogen)) or on ECM Gel (from Sigma) in MEF-conditioned KSR-hESC medium, at 37°C with 5 per cent CO_2. Subculture was performed on Day 7 using manual passaging of clones.

2.3 *Derivation of iPSCs*

Reprogramming of fibroblasts into iPSCs was carried out as described by Yu *et al.* [45] using a total of 10µg of total plasmid from two different combinations of episomal vectors that expressed *OCT4, SOX2, NANOG, LIN28, c-MYC, KLF4,* and *SV40LT* (plasmids were a gift from James Thomson, Combination 4; plasmid#20924: pEP4 E02S CK2M EN2L (7.3µg) and plasmid#20927: pEP4 E02S ET2K (Figure 2-2.A), and Combination 19; plasmid#20925: pEP4 E02S EN2K (3µg), plasmid#20927:pEP4 E02S ET2K (3µg) and plasmid#20926: pCEP4-M2L (2µg) (Figure 2-2.B) [45]. Clones were picked between Days 25-30 after transfection with the plasmids, screened for TRA-1-60 immunoreactivity and expanded for further analysis. After derivation, pluripotent cells were either maintained on MEFs in KSR-hESC culture medium or on ECM Gel in MEF-conditioned KSR-hESC medium as described above.

2.4 *Immunofluorescence assays*

The antibodies used were OCT4 (Millipore, MAB4401), NANOG (Cell Signalling, D73G4) TRA-1-60 (Millipore, MAB4360), TRA-1-80 (Millipore, MAB4381), smooth muscle actin (Millipore, CBL171), alpha-fetoprotein (Millipore, 2004189) and nestin (Millipore, ABD69). Cells were either grown on chamber slides or coverslips. Cells were fixed in 4 per cent paraformaldehyde for 15 minutes at room temperature before blocking with 5 per cent goat serum in 1X PBS. For intracellular or nuclear proteins, 0.2 per cent Triton X-100 was first used to permeabilise the cells at room temperature for 10 minutes before blocking with 5 per cent goat serum in 1X PBS. The blocked cells were incubated overnight at 4^0C with the primary antibody. This was followed by incubation with the appropriate secondary antibody at a 1:1,000 dilution. Slides were mounted with ProLong® Gold antifade mountant with DAPI (4',6-diamidino-2-phenylindole) to enable visualisation of the nuclei. An Olympus IX51 fluorescence microscope (Olympus, Tokyo, Japan, http://www.olympus-global.com) and Q-Capture Pro v6.0 software were used for image analysis and documentation.

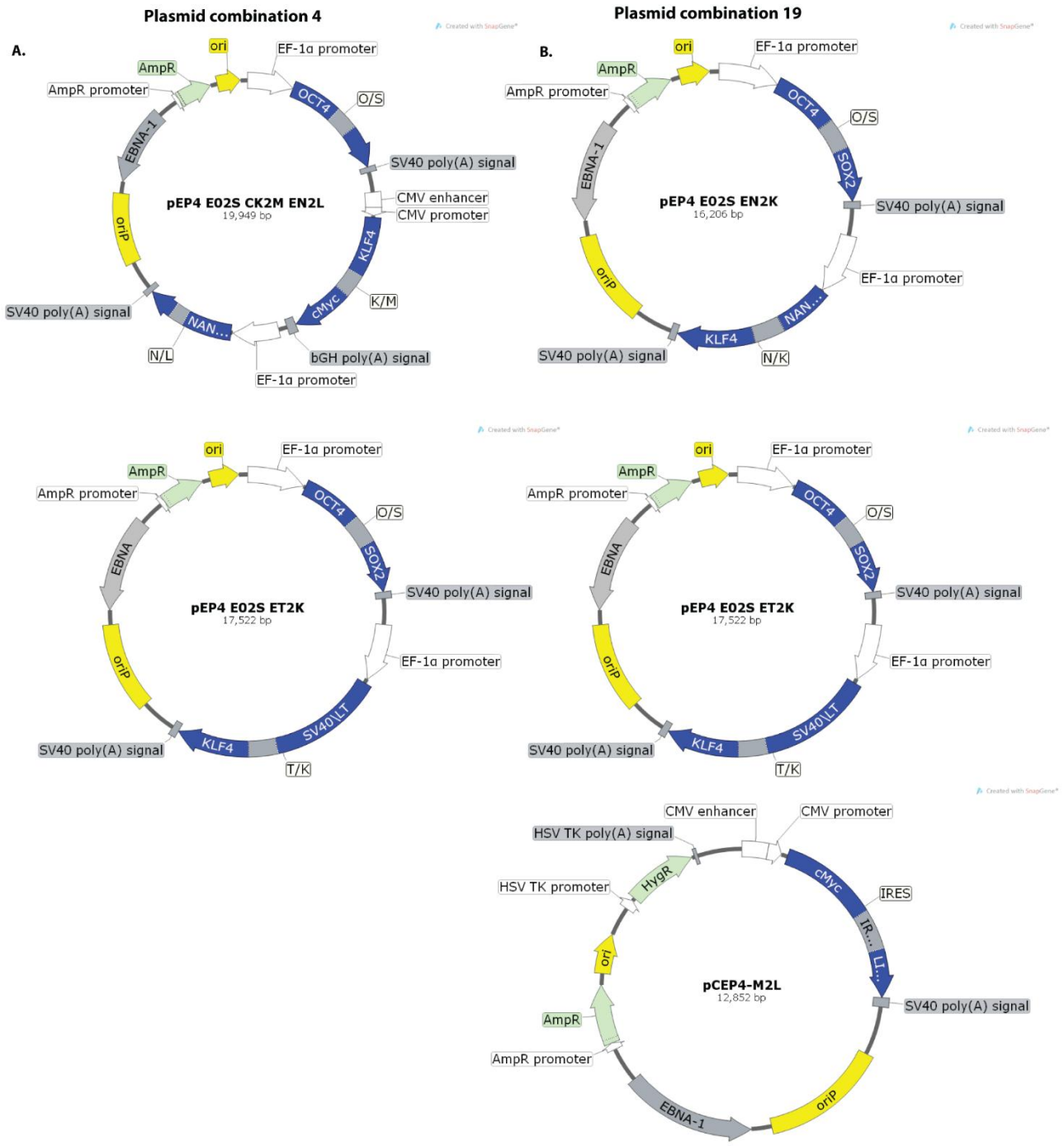

Figure 2-2: Schematic diagram illustrating the episomal plasmids in Combination 4 or Combination 19 used in fibroblasts reprogramming. (A) Combination 4: OSKMNL and OSTK-expressing plasmid. (B) Combination 19: OSTK, OSNK, and ML-expressing plasmid.

2.5 *Karyotype analysis*

Karyotype analysis (using standard G-banding) was carried out by Sullivan-Nicolaides Pathology (Taringa, Queensland, Australia, http://www.snp.com.au) to assess structural and numerical chromosomal abnormalities.

2.6 *Functional assays of pluripotency*

Standard *in vivo* (teratoma) and *in vitro* (embryoid body, EB) assays were used to assess the ability of iPSCs to differentiate into all three germ lineages. To generate EBs, mechanical dissociation was used to harvest clumps of cells. The clumps were suspended in hESC medium without bFGF on a low-attachment plate, to allow the cells to aggregate into EBs, and cultured for 20 days. The medium was changed every other day. Subsequently, EBs were either lysed in TRIzol for RNA extraction or dissociated with TrypLE and transferred to gelatin-coated wells for immunostaining. For teratoma assays, iPSCs were grown on MEFs and sent to StemCore (StemCore, Australian Stem Cell Centre, St Lucia, Queensland, Australia) for injection into NOD/SCID mice as per standard protocols. Injections were performed in duplicate for each iPSC clone. After eight weeks, teratomas were paraffin-embedded and sections stained with hematoxylin and eosin and examined for the presence of tissues representing the three germ layers.

2.7 *Vector construction*

2.7.1 The RGN plasmid

Single expression vectors for both a chimeric single gRNA driven by the U6 promoter and a human codon-optimised Cas9 driven by the CBh promoter [Cas9WT: pSpCas9(BB)-2A-GFP (PX458) (Addgene plasmid # 48138) or pSpCas9(BB)-2A-Puro (PX459) (Addgene plasmid # 48139) and Cas9n: pSpCas9n(BB)-2A-GFP (PX461) (Addgene plasmid # 48140) or pSpCas9n(BB)-2A-Puro (PX462) (Addgene plasmid # 48141, gift from Feng Zhang)] were used [177]. Guide RNA-expressing vectors were constructed as outlined in Ran *et al.* [177], with slight modifications. Briefly, gRNA sequences with a sequence pattern $GN_{19}NGG$ encoding the targeted sequences (Table 2-1) were designed using the online CRISPR design tool (http://tools.genome-engineering.org) [58]. A guanine residue was appended at the beginning of the gRNA sequence when the genomic sequence did not contain one. Complementary oligonucleotide pairs were ordered from Sigma. The annealed oligos formed overhangs compatible with ligation into the *Bbs*I-digested CRISPR-Cas9 plasmid backbone. The ligated vectors were transformed into competent cells and plated on agar plates containing carbenicillin (50 μg/ml). Three clones were picked, inoculated into LB broth with 50 μg/ml carbenicillin and incubated overnight at 37C on an orbital

shaker. The plasmid was extracted using the NucleoBond Xtra EF Kit *(*Macherey-Nagel). The presence of the insert was detected by PCR screening using the U6 forward primer and the complementary strand of sgRNA as a reverse primer. The correct insert sequence was further verified by Sanger sequencing.

Table 2-1: Target sequences for CRISPR/Cas9. RNA guide sequences used either to induce DNA strand break or to activate the targeted gene.

	Gene target	Location	gRNA	Guide sequence	GC%	G
DNA break: Cas9 /Cas9n	*HBB*	ExonII/ IntronII	1	AGCTCACTCAGTGTGGCAA	52.6	5
			3	GAACTTCAGGGTGAGTCTA	47.4	6
			4	AACTTAACCATAGAAAAGA	26.3	2
			5	TAGGAAGGGGATAAGTAAC	42.1	7
			6	AACAGGGTACAGTTTAGAA	36.8	5
			7	ACTTAACCATAGAAAAGAA	26.3	2
			8	TAAGTTCATGTCATAGGAA	31.6	4
			9	ACGAATGATTGCATCAGTG	42.1	5

CRISPRa	Gene target	Location	gRNA	Guide sequence	Position relative to TSS	
Gene activation: dCas9-VP16(n)	*HBG*1/*HBG*2	Proximal promoter	1	GCTAGGGATGAAGAATAAA	-26	*
			2	TTGACCAATAGCCTTGACA	-101	*
			3	TGCAAATATCTGTCTGAAA	-163	*
			4	AAATTAGCAGTATCCTCTT	-209	*
	HBB		1	GGTTGGCCAATCTACTCCC	-64	
			2	CCTGTGGAGCCACACCCTA	-84	
			3	CCAAGAGATATATCTTAGA	-189	
			4	AGGGCTGGGCATAAAAGTC	-23	
			5	CCAATCTACTCCCAGGAGC	-58	
			6	TCACTTAGACCTCACCCTG	-99	
			7	CAATTTGTACTGATGGTAT	-212	
			8	CCCTGTGGAGCCACACCCT	-85	

* Perez-Pinera et al. [79]

2.7.2 The CRISPRa plasmid

Gene activation was carried with vectors that simultaneously expressed the sgRNA and an inactive nuclease (dCas9) fused to a tandem repeat of the transactivation domain (TD) of the virion protein (*herpes simplex* virus) VP16 (four in VP64 and 10 in VP160). Cloning of the sgRNAs (Table 2-1) into the vectors was performed as described in the previous section, except that carbenicillin was replaced with kanamycin selection (at 50 μg/ml) for plasmid #48226. Plasmids used for transient transfections included pAC152-dual-dCas9VP64-sgExpression (sgRNA-dCas9VP64; Addgene plasmid #48238), pAC154-dual-dCas9VP160-sgExpression, (sgRNA-dCas9VP160 Addgene plasmid #48240), and pAC94-pmax-dCas9VP160-2A-puro (Addgene plasmid #48226). Plasmids were a gift from Prof. Rudolf Jaenisch [164]. (Vector#48226 was used during optimisation as detailed in the Results section).

2.7.3 The donor vector

To construct the donor vector for targeted correction, a 2,619 bp insert carrying a PGKpuroΔtk double selection cassette flanked by the ITR fragment was excised from the pMCS-AAT-*PB:*PGKpuroΔtk plasmid (kindly provided by Dr. Allan Bradley from Wellcome Trust Sanger

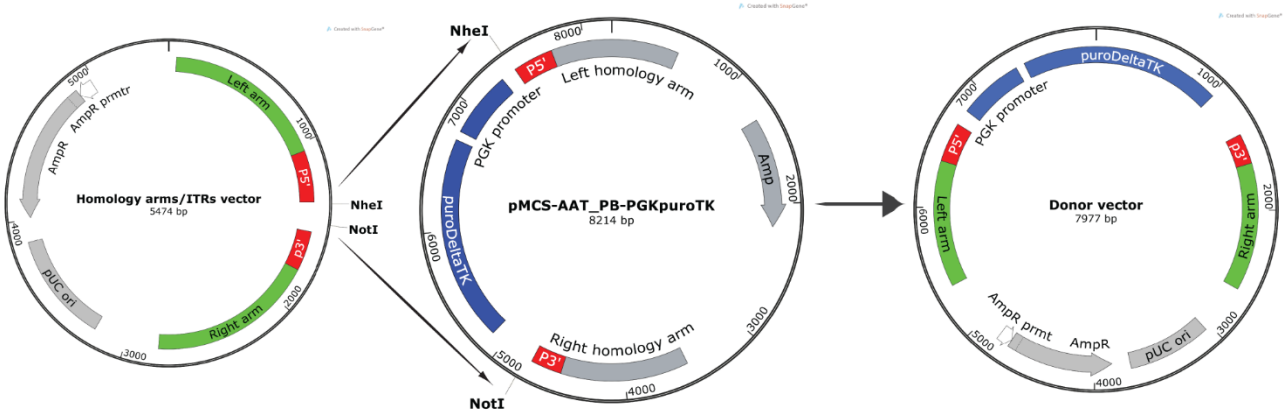

Figure 2-3: Schematic for cloning the PGKpuro-Δtk double selection cassette into the synthesised vector with the 5' and 3' homology arms.

Institute, Cambridge, UK) [178] using *NotI* and *NheI,* and cloned into a digested donor plasmid (synthesised by GenScript USA Inc, http://www.genscript.com) using the same enzymes (Figure 2-3). The fragment from the donor plasmid carried the 5′ and 3′ homology arms (approximately 1 kb each) and a segment of PB inverted terminal repeats to complement the fragment isolated from the pMCS-AAT-*PB:*PGKpuroΔtk plasmid. Isogenic sequences (to the parental beta thalassaemia-iPSCs) were used to synthesise homology arms (Genscript). Primers used to sequence the *HBB* locus are shown in Table 2-6.Three base substitutions and two SNPs, as well as the substitution of the disease-causing mutation, were introduced in the donor template sequence (Figure 4-6). See Section 4.2.3.1 for further details on the introduced SNP.

2.8 *Plasmid propagation, purification and quantification*

Ligated plasmid was transformed into chemically competent bacterial cells according to manufacturer's instruction (DHα: Invitrogen or Bioblue: Bioline) and plated on agar plates containing the appropriate antibiotic. Clones were picked after 24 hrs, inoculated into 100 ml LB broth medium supplemented with the appropriate antibiotic and incubated overnight at 37C on an orbital shaker (200rpm). The plasmid was extracted using the NucleoBond Xtra EF Kit (*Macherey-Nagel).* The plasmid was eluted using a minimum volume of elution buffer to obtain a final concentration of >1µg/µl. The plasmid yield and 260/280 ratios were quantified using a NanoDrop

1000 spectrophotometer (NanoDrop, Wilmington, DE, http://www.nanodrop.com). The presence of the insert was confirmed by sequencing.

2.9 *Transfections of CRISPR/Cas9 plasmid*

Various method are currently available to introduce DNA plasmid into cells. The method of choice depend on the cell type and the objective of the transfection [179]. To deliver plasmid expressing CRISPR components into cells, nucleofection or lipid-based transfection are commonly used [78, 94, 177, 180]. However, optimization of experimental conditions for both nucleofection and lipofection are needed in order to balance cell viability and efficiency of transfection. Optimization are further discussed in chapter 4 and 5

2.9.1 RGN

To induce DNA breaks in HEK293FT, the cells were cultured as previously described and seeded into a 12-well plate (about 130,000/well) one day prior to transfection. Unless stated otherwise, 1 μg of the dual sgRNA-Cas9-expressing vector or 500 ng of each of the pair of nickase plasmids was transiently transfected into 293FT cells using Lipofectamine LTX with the PLUS Reagent (Invitrogen) according to the manufacturer's instructions. Medium was replaced after 24 hours.

Nucleofection of iPSCs was carried out as described in Yang *et al.* [78]. Briefly, before nucleofection, cells were treated with 10 μM Y-27632 ROCK inhibitor for at least two hours. 2x 10^6 iPSCs were collected by trypsinisation and resuspended in 20 μl of Nucleofector mix that included P3 and supplement solutions (4D Nucleofector, Lonza, http://www.lonza.com). The mix was transferred to one well of a nucleovette strip that contained an equivalent of 1.5-2 μg of the RGN expression plasmid or an equal mixture of plasmids for paired nickase (in a maximum volume of 2 μl). The cells were nucleofected using the CB150 programme. For HDR, 500 ng of the donor vector was added to the suspension mixture. Nucleofected cells were plated in one well of a 6-well plate on feeders in KSR-hESC medium supplemented with 10 μM Y-27632 ROCK. ROCK inhibitor was removed from the medium after 24 hours. Puromycin selection was started after 72 hours post-transfection, using 1 μg/ml for three days, then maintained at 0.5 μg/ml for a minimum of 10 Days.

2.9.2 CRISPRa

HEK 293FT cells were cultured and transfected as previously described. Optimization of transfection conditions was performed as described in the Results section. In order to optimise transfection conditions, parameters including the total pDNA (amount of dCas9+gRNA), the amount of vectors carrying each sgRNA, and the charge ratio (Lipofectamine/DNA) were tested for

HBG using a previously published gRNA combinations [79]. Subsequently, optimised conditions were used to screen for the sgRNA mixture that displayed most robust activation of the *HBB* gene in HEK293FT. Multiple combinations were selected for further analysis in iPSCs. Nucleofection of iPSCs was carried using a P3 Primary Cell 4D-Nucleofector X Kit S (for 4D Nucleofector, Lonza, http://www.lonza.com). A total amount of 1.2-2 µg of dCas9-based activation vectors was used. This amount was equally divided between multiple vectors targeting different sites in the proximal promoter region of the *HBB* gene.

2.10 *Testing the functionality of single or paired guide RNAs in HEK293FT cells using the Surveyor assay*

Cells transfected with the specified vector or combination of vectors were collected by trypsinisation after 72 hours and the Surveyor assay was carried as descried by Ran *et al.* [177]. Briefly, DNA was extracted and amplicons that spanned the guide binding site were generated by PCR using various Surveyor primer pairs (Table 2-2).

Table 2-2: Primer sequences used for Surveyor assays

Primer name		Sequence	Amplicon size
S.V1	F	GAGACTTCCACACTGATGCAAT	658
	R	AAGTCAGGGCAGAGCCATCTAT	
S.V2	F	AGTAATGTACTAGGCAGACTGTG	616
	R	TGGGTTTCTGATAGGCACTGAC	
S.V3	F	AAAGCAGAATGGTAGCTGGATTGT	969
	R	TACCCTTGGACCCAGAGGTTCT	

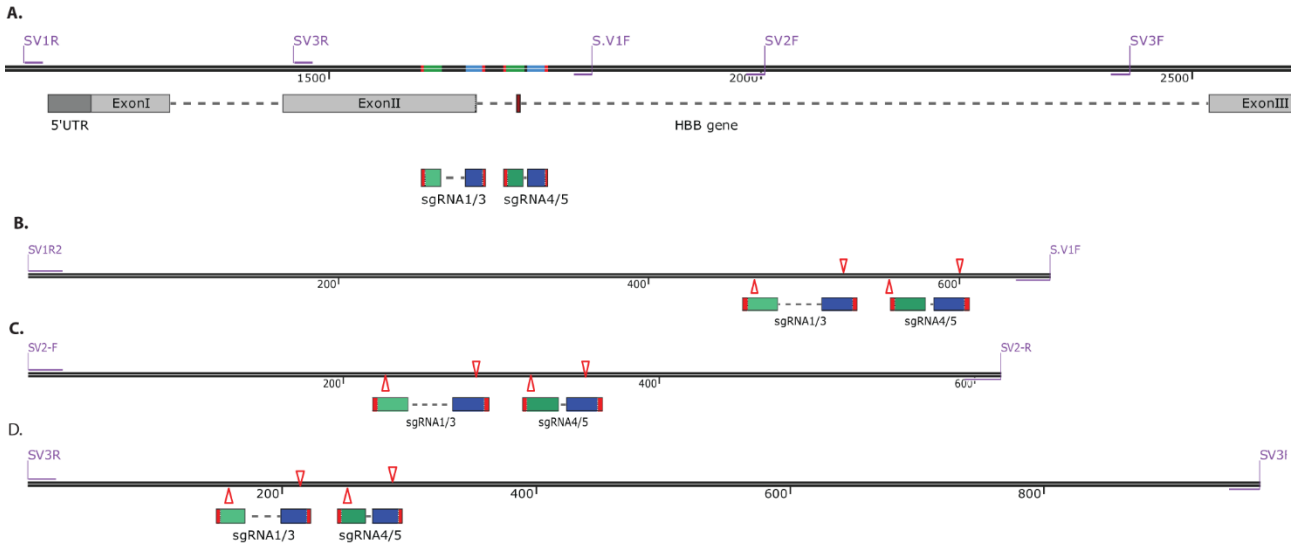

Figure 2-4: Schematics illustrating the location of the primer binding sites and Cas9 cleavage sites for various PCR-amplified products used in Surveyor assay. (A) The binding site for SV1, SV2, and SV3 primers.(B, C, and D) The predicted cleavage sites (red triangles) for sgRNA1/3 and sgRNA4/5 in PCR fragments generated by using primer pairs (B) SV1, (C) SV2, and (D) SV3 that span the targeted site in the β-globin gene.

The amplicons were purified using the QIAQuick PCR purification kit (Qiagen). Heteroduplexes were formed by subjecting the purified amplicons to heat treatment following by gradual cooling to 4^{o}C. The normalised solutions were treated with Surveyor nuclease according to manufacturer's instructions. Bands were resolved using 2 per cent agarose gels. Relative band intensity (volume intensity) was quantified using Image lab software (Bio-Rad) after subtracting background and adjusting the band limits to the band edges [181]. The percentage of indel mutation induced by Cas9 was calculated as indicated in Ran *et al.* [177] using the following formula:

Indel (%)= 100 x $[1-(1-f_{cut})^{1/2}]$ where $f_{cut} = \frac{\text{(Sum of band intensity of digested products)}}{\text{(Sum of band intensity of digested + undigested products)}}$

2.11 *RT-PCR and qRT-PCR expression analyses*

For RNA extraction, cells were lysed in TRIzol reagent (Ambion) followed by purification using the PureLink RNA Mini Kit (Ambion). One microgram of RNA was used for reverse transcription with the iScript cDNA synthesis kit (Bio-Rad, Hercules, CA, http://www.bio-rad.com) according to the manufacturer's specifications. Amplicons were generated from cDNA samples using PCR conditions specified in the next sections. Quantitative RT-PCR (q-PCR) analysis was performed in triplicate (or duplicate) using the CFX96 real-time PCR detection system (Bio-Rad) with Ssofast evagreen q-PCR mix (Bio-Rad). More than one housekeeping gene was used to normalise gene expression: *GAPDH, EEA1F1* (eukaryotic translation elongation factor1 alpha1) and *ACTB* (beta-actin). Expression data were either reported as relative gene expression (ΔCt) or fold change relative to a control (ΔΔCt). The former (ΔCt) was used when no expression of the gene of interest was detected in the control sample (for 40 cycles) in order to identify the combination of sgRNA yielding the highest increase in transcript abundance following activation. Sets of primers used to amplify the intended locus are listed in Table 2-6.

2.12 *PCR amplification and sequence verification*

Genomic DNA was extracted using the NucleoSpin tissue kit (Macherey-Nagel), or ISOLATE II genomic DNA kit from Bioline according to the manufacturer's instructions, and episomal DNA was isolated using the QIAprep spin miniprep kit (Qiagen), according to the manufacturer's instructions. PCR reactions were performed using either Phusion high-fidelity DNA polymerase (New England Biolabs) for sequencing of the *HBB* locus, Herculase II fusion polymerase (Agilent Technologies) for sequencing of the targeted clones and for the Surveyor assay (Surveyor Mutation

Detection Kits), LongAmp Taq DNA polymerase (New England Biolabs) for the PCR-based screening for the targeted integration of the selection cassette, or Scientifix Taq DNA polymerase (TDP-500) for screening of the reintegration of the selection cassette. Touchdown PCR was mostly used to amplify the specified locus (25-30 cycles) according to the manufacturer's instructions (depending on the Taq polymerase used). Sanger sequencing was used to verify the sequences for some samples and plasmids. Sequencing was performed by the Australian Genome Research Facility (AGRF, Brisbane node). Primer sequences are shown in Table 2-6.

2.12.1 PCR-based screening of modifications in iPSC clones

2.12.1.1 *Genotyping of puromycin-resistant clones*

Genotyping of puromycin-resistant clones was carried out as described by Yusa [182]. Briefly, cell lysate was prepared by cutting a small section from clones, suspending it in water with subsequent heat treatment for 10 minutes. This was followed by incubation for 1 hour with proteinase K. Cell lysate was used in multiplex PCR (Figure 2-5) in the presence of two sets of primers: one for genotyping the right arm and the other for the left arm (Table 2-3). Amplification was carried out using LongAmp Taq DNA polymerase as detailed in the Yusa's [182] protocol. PCR amplicons were analyzed on a 1.5% agarose gel. Band pattern was used to screen for the absence or the presence of a possible targeting event at one or both alleles. DNA from the parental iPSC clone was used a positive control to screen for the unmodified original allele and also served as a negative control for the targeted allele.

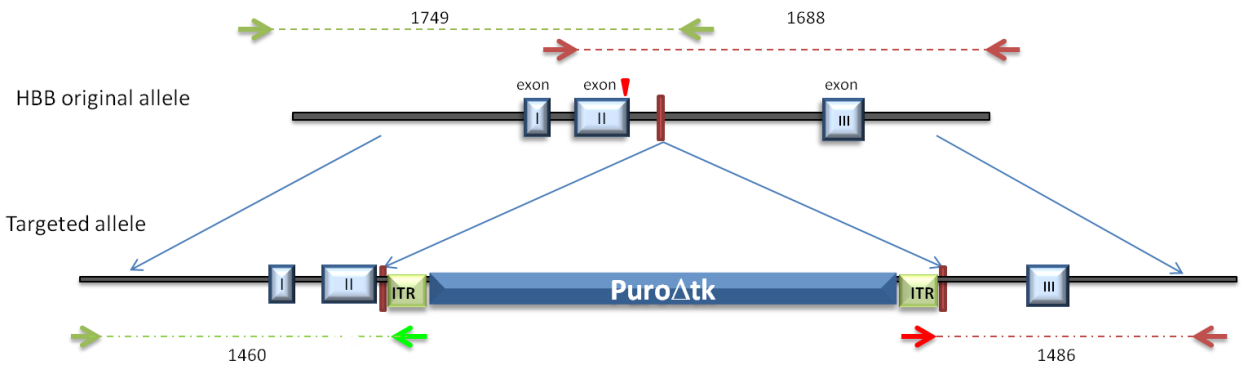

Figure 2-5: The genotyping strategy used to screen for the presence of the targeted allele and/or the original alleles in puromycin-resistant clones. Green and red arrows indicate primer sets used to screen for the presence of the original allele or targeted modification on the left arm and right arm, respectively.

Table 2-3: Primer sequences used for PCR-based screening of the left and right arms in resistant clones

Primer name	Sequence		
F2	TAGTAATGTACTAGGCAGACTGTG		Left arm
R2	TTGCAGATTAGTCCAGGCAGAAA		
P5*	GCGACGGATTCGCGCTATTTAGAAAG		
F1	AAGGCTGAGAGATGCAGGATAAG		Right arm
R1	GTGGAGACAGAGAAGACTCTTGG		
P3*	CGTCAATTTTACGCATGATTATCTTTAAC		
Multiplex PCR	Targeted allele	Original allele	
F1R1P3	1460 bp	1749 bp	
F2R2P5	1486 bp	1688 bp	

*Primers reported in Yusa *et al.* [182]

2.12.1.2 *Screening for random integration*

Screening for on-target and random integration was performed simultaneously on the same cell lysate using the same PCR conditions described above. The sets of primers used are listed in Table 2-4. Samples with positive bands at either arm were excluded from further analysis.

Table 2-4: Primer sequences used in PCR-based screening for random integration of the donor vector

Primer name	Sequence	DNA region	Expected band size
M13F97*	TCGGTGCGGGCCTCTTCGCTATTAC	Left arm	1,141 bp
T3.1	GCGTCCCATAGACTCACCCTGAA		
M13R84*	ACTCATTAGGCACCCCAGGCTTTAC	Right arm	1,892 bp
PurF*	CTGCTGCAACTTACCTCCGGGATG		

*Primers reported in Yusa *et al.* [182]

2.12.1.3 *Heteroduplex mobility assays*

Heteroduplex mobility assays (HMA) were used for allelic alteration screening at on- and off-target sites. The screen was carried out in puromycin-resistant clones with HR as described by Zu *et al.* [183], with minor modifications. To screen for on-target mutations, genomic DNA was extracted from both parental and puromycin-resistant iPSC clones at the same passage number. PCR amplicons for the DNA region flanking the Cas9 targeting sites were amplified from iPSC clones (resistant and parental clones) and the donor vector using three sets of primers. The first set was used to detect indel in the non-targeted allele, and the other two were used to screen the left and the right arms of the targeted allele (Figure 2-6). The primers sequences used are listed in Table 2-5.

Table 2-5: Primer sequences used to screen for the presence of allelic modifications in HMA

	Primer name	Sequence	DNA region for indel screening	Expected band size
A.	MA233R	GTGCTCGGTGCCTTTAGTGAT	Original allele	233 bp
	MA233F	GCAATCATTCGTCTGTTTCCCATT		
B.	P3	CGTCAATTTTACGCATGATTATCTTTAAC	Targeted allele (right arm)	130 bp
	MA233F	GCAATCATTCGTCTGTTTCCCATT		
C.	P5	GCGACGGATTCGCGCTATTTAGAAAG	Targeted allele (left arm)	249 bp
	MA233R	GTGCTCGGTGCCTTTAGTGAT		

PCR products were assessed for the correct size on 2 per cent agarose gel. In order to generate heteroduplexes, PCR products from the control sample (BT-iPSCs or donor vector) were mixed with products from the tested sample at a ratio of 1:1 and subjected to denaturing and reannealing procedures. The annealed products were analysed by electrophoresis using 12 per cent native polyacrylamide gels. The gel mixture was prepared using acrylamide-bisacrylamide (29:1, w/w), 1x Tris-borate-EDTA (TBE), ammonium persulfate, and tetramethylethylenediamine (TEMED). Samples were loaded into wells and subjected to a constant voltage of 120V/cm for ≈2 hours. The gel was immersed in diluted SYBER Gold solution for 35 minutes before visualisation using the Geldoc XR1 Imaging System (Biorad). The allelic alteration was identified by comparing samples to a control homoduplexes. The genotype of selected samples was further validated with Sanger sequencing.

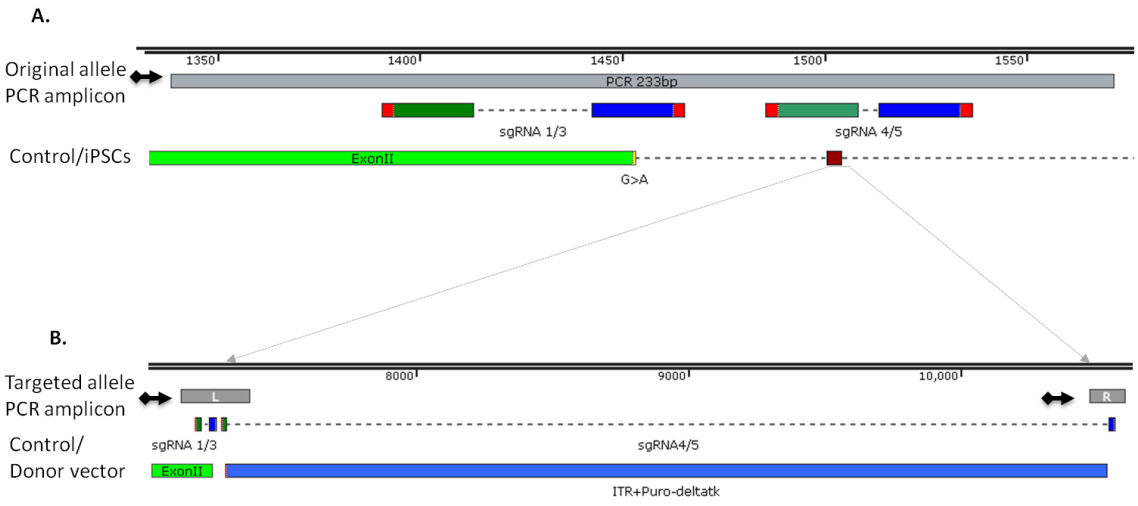

Figure 2-6: Strategy for screening for allelic alteration at specific DNA regions using HMA. A) PCR amplicons used to screen for possible DNA alterations introduced by the Cas9 nickase at the on-target locus in the original allele compared to that from BT-iPSCs. B) Two PCR amplicons were used to screen for the targeted allele compared to that from donor vector. ↔ indicates PCR amplicons with their relative position to the paired nickase used to stimulate the gene targeting event.

For off-target screens, the same steps as described in the previous section were followed using the primer listed in Table 2-6.

2.13 *Excision of the selection cassette*

Excision of the selection cassette was carried by transfecting targeted iPSCs with a 2 µg of a plasmid expressing the HA-tagged hyperactive *piggyBac* transposase (kindly provided by Dr. Allan Bradley) using the above-mentioned transfection conditions in iPSCs [184]. FIAU selection was ineffective, as demonstrated by the high survival of a negative control (biallelic targeted clones). A stepwise enrichment procedure, as shown in Figure 2-7 was adopted. This procedure is detailed in the Result section of Chapter IV. Multiplex PCR-screening was carried using the primers listed in Table 2-5A. and B to screen for the restoration of the TTAA site and the presence of the targeted selection cassette, respectively.

Absence of reintegration was confirmed in the selected clones using a PCR-based method with a primer pair located inside the puromycin selection cassette.

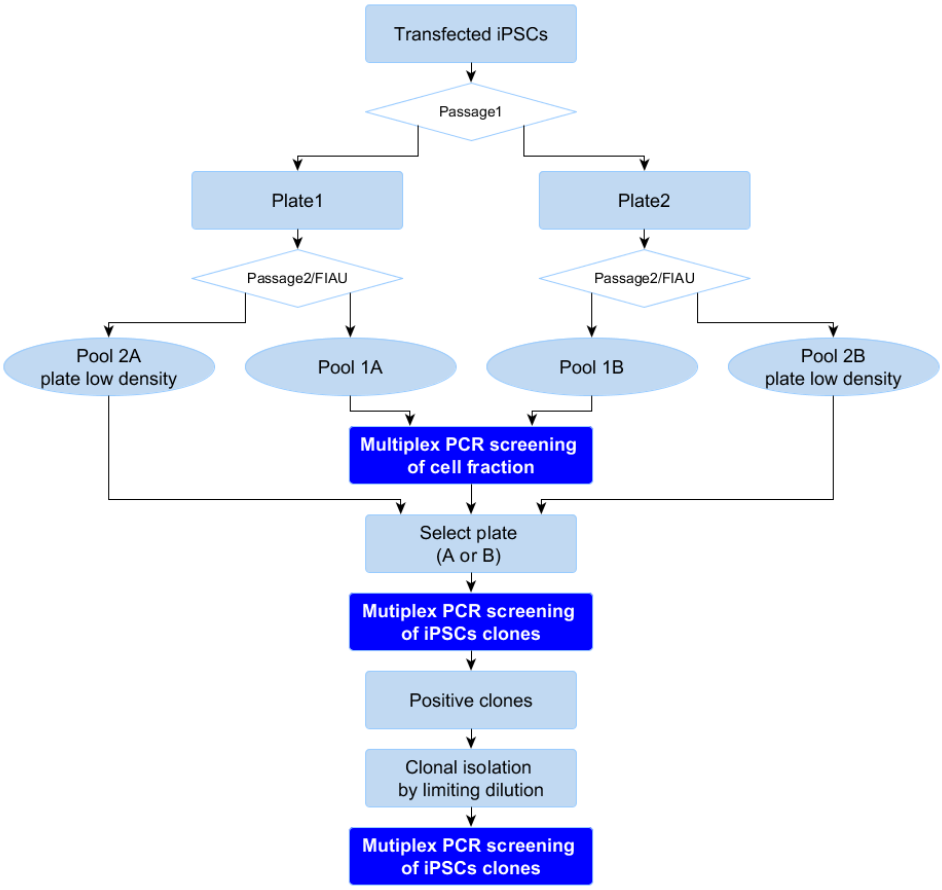

Figure 2-7: Outline of the strategy used to enrich and isolate iPSC clones with a restored TTAA site after excision of the cassette.

Table 2-6: Primer sequences

	Primer name		Sequence	band size bp	
Transgene					
	EBNA	F	TTTTCGCTGCTTGTCCTTTT	883	PCR/RT-PCR
		R	TTCCAACCCGAAATTTGAGA		
	Lin28	IRES2-SF	TGGCTCTCCTCAAGCGTATT	245	PCR/RT-PCR
		Lin28-SR	GCAAACTGCTGGTTGGACAC		
	KLF4	IRES2-SF	TGGCTCTCCTCAAGCGTATT	253	PCR/RT-PCR
		Klf4-SR	GTGGAGAAAGATGGGAGCAG		
	OCT4	Oct4-SF1	AGTGAGAGGCAACCTGGAGA	657	PCR/RT-PCR
		IRES2-SR	AGGAACTGCTTCCTTCACGA		
	NANOG (exo)	Nanog-SF2	CAGAAGGCCTCAGCACCTAC	732	PCR/RT-PCR
		IRES2-SR	AGGAACTGCTTCCTTCACGA		
	GAPDH	F	GTGTGGCTCCCTTGGGTATATG	351	PCR/RT-PCR
		R	ATACTTCTCATGGTTCACACCCA		
	NANOG (endo)	F	TTTGGAAGCTGCTGGGGAAG	194	PCR/RT-PCR
		R	GATGGGAGGAGGGGAGAGGA		
Selection cassette	PURO	F	CTGCTGCAACTTACCTCCGGGATG	323	PCR
		R	CCAATCCTCCCCCTTGCTGTCCTG		
EB	GATA4	F	GCGGTGCTTCCAGCAACTCCA	138	q-PCR
		R	GACATCGCACTGACTGAGAACG		
	MAP2	F	CAGGTGGCGGACGTGTGAAAATTGAGAGTG	186	q-PCR
		R	CACGCTGGATCTGCCTGGGGACTGTG		
	PAX6	F	CAGCACCAGTGTCTACCAACCA	67	q-PCR
		R	CAGATGTGAAGGAGGAAACCG		
	SMA	F	ACGTGGGTGACGAAGCACAG	135	q-PCR
		R	GGGGAACACGAAGCTCATTGT		
	Brachyury-T	F	TCAGCAAAGTCAAGCTCACCA	102	q-PCR
		R	CCCCAACTCTCACTATGTGGATT		
	GAPDH (exon)	F	GTGGACCTGACCTGCCGTCT	152	q-PCR
		R	GGAGGAGTGGGTGTCGCTGT		
	SOX17	F	GGCGCAGCAGAATCCAGA	61	q-PCR
		R	CCACGACTTGCCCAGCAT		
	SOX2	F	CCACCTACAGCATGTCCTACTCG	123	q-PCR
		R	GGGAGGAAGAGGTAACCACAGG		
	OCT4	F	CGTGAAGCTGGAGAAGGAGA	91	q-PCR
		R	CTTGGCAAATTGCTCGAGTT		
	GAPDH	F	ACCCACTCCTCCACCTTTGAC	100	q-PCR
		R	TGTTGCTGTAGCCAAATTCGTT		
	EEF1A1	F	CCCCAGGACACAGAGACTTT	120	q-PCR
		R	GCCCATTCTTGGAGATACCA		
	B-Actin	F	GCCGGGGACCTGACTGACTAC	101	q-PCR
		R	TTCTCCTTAATGTCACGCACGAT		

Table 2-6 Continued

Primer name		Sequence	band size bp	
HBD	F	ATCCTGAGAACTTCAGGCTCTT	161	q-PCR
	R	AAACAGTCCAGGATCTCAATGGT		
HBG	F	GGACAAGGCTACTATCACAAGC	193	q-PCR
	R	GGAAGTCAGCACCTTCTTGC		
*HBB*1	F	GGCACCTTTGCCACACTG	132	q-PCR
	R	CACTGGTGGGGTGAATTCTT		
*HBB*S	F	TGCACGTGGATCCTGAGAACT	79	q-PCR
	R	AATTCTTTGCCAAAGTGATGGG		
*HBB*3	F	CTCATGGCAAGAAAGTGCTCG	181	q-PCR
	R	AATTCTTTGCCAAAGTGATGGG		
NANOG	F	TTTGGAAGCTGCTGGGGAAG	194	q-PCR
	R	GATGGGAGGAGGGGAGAGGA		

Off-target analysis	Primer name		Sequence	Band size bp	
sgRNA1	G1-11-S1F	F	ACTCTCAAAATCCCACTGCTGAA	365	PCR
		R	AGTCTCCTGGTTGTCTACCCAT		
	G1-8-S2F	F	ATCTCTGAGAAGCTCAAGGCTC	335	PCR
		R	AGAACTGAAGGCAGAAGAAGTGT		
	G1-4-S3F	F	GTGTCTGCCTAGACTGTGTTAGA	370	PCR
		R	AGGATCTCAATAGCCCTGAGTG		
	G1-1-S4F	F	CTGGAGCACGCTTTAGACCAG	333	PCR
		R	CCCTGAAATGGCTCCTTAGAGAC		
	G1-13-S5F	F	GAACACGAGTAGGAGTGTTGTG	374	PCR
		R	CCTGGCAAATGAACTGTCACT		
sgRNA3	G3-2-S1F	F	GGCGCCCTATACTCCATTCTTCA	314	PCR
		R	ACGGTTTCATGCTGGGATTCCATA		
	G3-8-S2F	F	GTACTTGGGCTCCGAGTTTCTG	342	PCR
		R	CACCTCCTTACTTGAAGCAGCA		
	G3-6-S3F	F	GTGTTTGGGTATTTACTCCCCCT	343	PCR
		R	AGGAAAACTGACACAGGAAATGC		
	G3-X-S4F	F	ATGTACATGTCCCACGGCTC	318	PCR
		R	TTCGGCTTGCTAAACACCCTG		
	G3-12-S5F	F	CTTGGCTTTCCTTGCCCTATAC	336	PCR
		R	TGCAATCCAGATTCCATCCTGA		
sgRNA4	G4-8-S1F	F	CCAATAATGCCCTTCTGGTCTG	330	PCR
		R	ATGCCAACTGTGTAACAAGGAC		
	G4-X-S2F	F	AAATTGCAATGCATGTCCTCACC	372	PCR
		R	CAGGTCGATTCTGGTGGCAAGAT		
	G4-2-S3F	F	GTTTGATTTTAGCCGCAGAAACTG	309	PCR
		R	CGTCATTAATATCTCACAGGCAGAA		
	G4-6-S4F	F	GTATGATTAGCTGAAGGGAGCTGG	337	PCR
		R	GCAAACCAAGCCACAGGAAACT		
	G4-12-S5F	F	TCAATTCTAGAGCACACAGGCT	307	PCR
		R	GCAATGGTCGATCTAAAGCTGC		
sgRNA5	G5-10-S1F	F	ACCTCAAGCCTTTTCCTGATTCT	327	PCR
		R	TGCTCTGCTGGGATGTAATGAAT		
	G5-2-S2F	F	AGAACCTAATTTCCAAGGGAGCC	339	PCR
		R	CCGAAGAGACTTTGGAATGAAAATC		
	G5-5-S3F	F	CCCTACATGTGACAGGATAGCTG	348	PCR
		R	AGCAGGCATTGCAAGAGATAAA		
	G5-14-S4F	F	CTTCCTGATGAGCCAGTAACCC	327	PCR
		R	TCTTCTCATCAAGAGCACAGGT		
	G5-2-S5F	F	AGTCCGTCTGCTGTAGGAACAC	304	PCR
		R	AACGAAAGGAGCACATCAGAGAT		

Table 2-6 Continued

Sequencing *HBB* locus					
***HBB* PCR primers**	A	F	ACGGCTGTCATCACTTAGAC	1457	PCR
	B	R	AAGAGGTATGAACATGATTAGC		
	C	F	GTGTACACATATTGACCAAATC	1213	PCR
	D	R	CAGATTCCGGGTCACTGTG		
	T3.9F	F	CCATCACTAAAGGCACCGAGCAC	1324	PCR
	T3.9R	R	AGACCTTTTCCCCTCCTACCCCTA		
Sequencing primers	T3-ABS-F	F	CAAAGGACTCAAAGAACCTCT		
	T3ABS-R	R	GAGAAGATATGCTTAGAACCGA		
	ABS F	F	TAATGTACTAGGCAGACTGTG		
	ABS R	R	CAGAGCCATCTATTGCTTACA		
	ABS2 F	F	GACTCAGAATAATCCAGCCTT		
	ABS2 R	R	CATGTCATAGGAAGGGGATAAG		
	CDS F	F	TTCATGAGGACAGCTAAAACAA		
	CDS R	R	GTGATAATTTCTGGGTTAAGGC		

CHAPTER III:

THE GENERATION OF TRANSGENE-FREE INDUCED PLURIPOTENT STEM CELLS WITH A BETA THALASSAEMIA MUTATION

3 THE GENERATION OF TRANSGENE-FREE INDUCED PLURIPOTENT STEM CELLS WITH A BETA THALASSAEMIA MUTATION

3.1 *Introduction*

Beta thalassaemia is an umbrella term for a group of inherited blood disorders that are characterised by an absence, reduction, or otherwise abnormal synthesis of the haemoglobin subunit beta (β-globin chain, *HBB* gene), mostly brought about by point mutations, or small deletions, in the β-globin gene causing defective transcription, splicing or translation of β-globin mRNA. Beta thalassaemia is a relatively frequent genetic disorder, with an estimated 5 percent of the world population being carriers [1]. Patients with beta thalassaemia major are homozygotes or compound heterozygotes for a severe type of *HBB* mutation and display severe anaemia and hepatosplenomegaly. They require regular blood transfusions as a supportive therapy, with the only currently available effective treatment being an allogenic HSC transplant. Given the incidence of the disease, and the shortage of suitable HLA-matched donors, engraftable autologous cells derived by *ex vivo* manipulation of the patient's HSCs, through the directed differentiation of gene-edited patient-specific iPSCs, or the direct conversion of gene-edited somatic cells into HSCs, offers an alternative *de novo* source of cells with the potential for clinical applications [24].

Although recent evidence has suggested that *ex vivo* targeted gene correction in HSCs is achievable [6], this approach limits the number of cells available for gene therapy due to the inefficiency of gene targeting in these cells and their poor expandability *in vitro*; thus, a successful therapeutic effect can only be achieved through introduction of a gene that confers a survival advantage [5, 8]. Conversely, iPSCs present a practical alternative that allows careful verification, enrichment and expansion of gene-edited cells with subsequent haematopoietic differentiation capacity. Provided engraftable HSCs can be generated from iPSCs, this approach may allow for the generation of therapeutic levels of isogenic HSCs. However, to be clinically relevant, the genetic defect should be precisely corrected, preferably using non-integrating methodologies so as to maintain the integrity of the genome.

Several integration-free methods are available for the derivation of footprint-free iPSCs. Footprint-free methods can involve the use of a plasmid, minicircles, episomes, mRNA or the Sendai-virus to achieve transient ectopic expression of reprogramming factors [18]. Alternatively, cells can be transduced with the reprogramming factors in the protein form [42, 43, 185]. Episomal-based reprogramming of somatic cells has been reported to be a more reliable method for iPSC derivation from fibroblasts in comparison with mRNA, result in a rapid and passive loss of the transfected

vector in comparison to Sendai virus-based reprogramming, and more importantly, iPSCs may be more appropriate for potential therapeutic use [186].

Therefore, this study aimed to generate iPSCs from beta thalassaemia patients using episomal-based reprogramming. While the generation of iPSCs from the fibroblasts of beta thalassaemia individual [26-28, 35-37] and amniotic fluids cells [38] has previously been reported [28, 187] he production of footprint-free iPSCs from fibroblasts derived from patients with the Mediterranean mutation IVSII-1(G>A) of beta thalassaemia has not been reported to date (Table1-1). IVSII-1(G>A) is a splice junction mutation that affects the 5′ splice site leading to a severe form of beta thalassaemia in the homozygous state. Restoration of the normal splice site in gene-edited cells can be verified by assessing the formation of the normally spliced product (mRNA) generated upon *HBB* gene activation.

3.2 *Results and discussion*

3.2.1 Establishment of fibroblast cultures and derivation of iPSCs

All primary samples were collected from donors by a trained physician after obtaining informed consent. Punch biopsies were acquired from the inner arm of three patients homozygous for beta thalassaemia mutations: siblings with the IVSII-1 (G>A) mutation (a 9-year old female and a 10-year old male, referred to as T3 and T2, respectively) and one patient with the IVSI-110 (G>A) mutation (16-year old male, T1), as well as from a control individual (29-year old male, N4). While IVSII-1 (G>A) results in the absence of β-globin production ($beta^0$ thalassaemia), the IVSI-110 (G>A) mutation leads to a substantial reduction in the synthesis of β-globin ($beta^+$ thalassaemia). The former mutation affects the donor splice site abolishing normal splicing at this location [188], while the latter introduces an alternative acceptor splice site that results in a significant reduction in the correctly spliced product [189]. Fibroblast cultures were established using the explant outgrowth method as described by Raya *et al.* [176] with minor modifications. At Days 4-5 of culture, cells with a morphology consistent with that of keratinocytes were observed in some explants (Figure 3-1.A, B, C and D). Fibroblasts emerged later and dominated the culture surface except in the small area where keratinocytes emerged from the skin specimen (Figure 3-1.C and D). Keratinocytes stopped proliferating, presumably due to unfavourable culture conditions, as the fibroblasts were cultured in a medium with a high calcium concentration and in the absence of keratinocyte growth factors. Keratinocytes became more phase-dense and remained on the culture plates after the fibroblasts were passaged with TrypLE**.** After passaging, the morphology of the cells was almost entirely consistent with fibroblasts (Figure 3-1.F). Primary fibroblast cultures were verified to be free of transmissible pathogens (including mycoplasma) and cryopreserved at Passage 2. The presence of the mutation was confirmed in the fibroblasts derived from the three beta thalassaemia patients using the β-globin strip assay MED kit (Vienna Lab) (Figure 3-1.G).

Next, fibroblasts were transfected with episomal plasmids as described by Yu *et al.* [45]. Transfection was carried-out using either Combination 4 (C4) or Combination 19 (C19). C4 is composed of two vectors: the first vector expressed *OCT4, SOX2, KLF4, C-MYC, NANOG,* and *LIN28* (OSKMNL) under either the strong elongation factor alpha (EFα) or cytomegalovirus (CMV) promoter, and the second expressed *OCT4, SOX2, SV40LT* and *KLF4* (OSTK) under the EF1α promoter, whereas C19 is composed of three vectors that express the same set of transgenes (OSTK, OSNK, ML) except for differences in their arrangement and the type of promoter used for some of the transgenes (Figure 2-2). Transfected cells were seeded onto irradiated MEF feeders and adapted from fibroblast medium to hESC medium between Days 4 and 6. From Day 9 onwards,

MEF-conditioned medium supplemented with 100 ng/ml b-FGF was used until around Day 25-30 (Figure 3-2.A) [190].

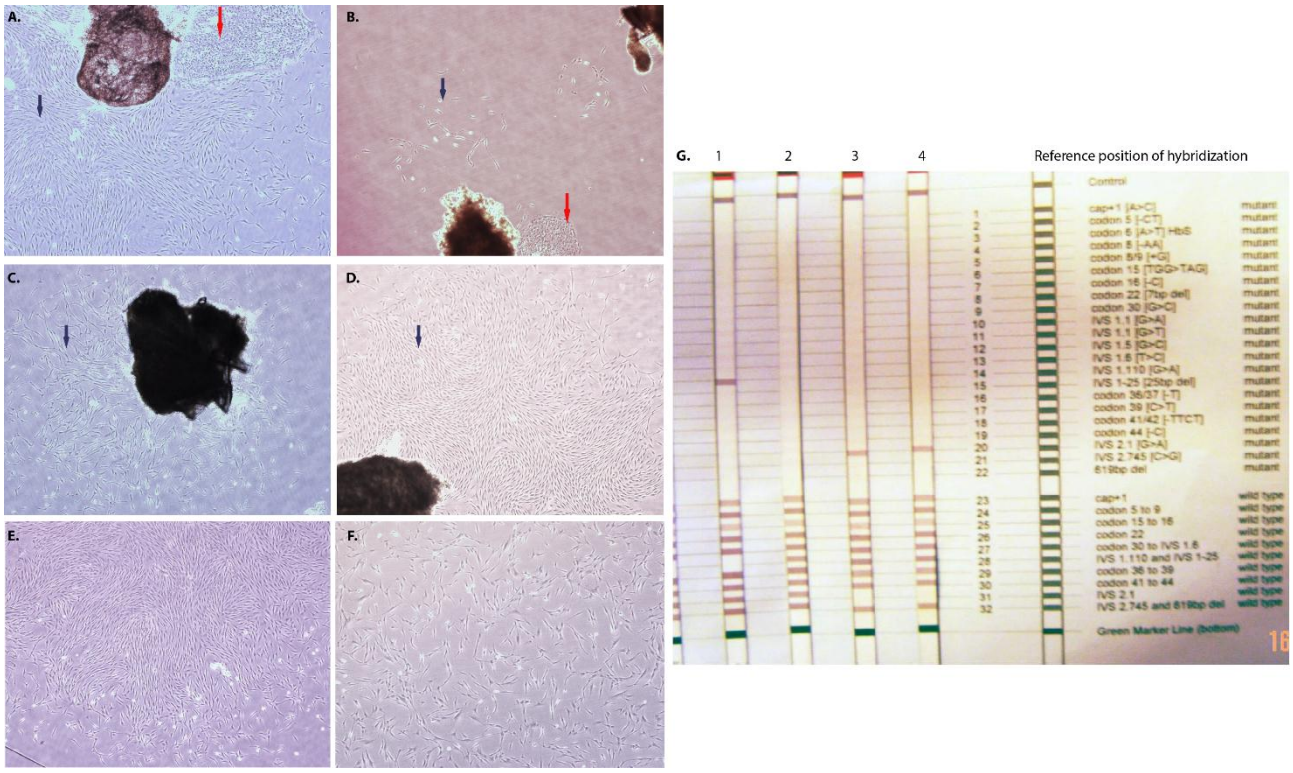

Figure 3-1: Establishment of fibroblast cultures and characterisation of the β-globin gene mutation using reverse hybridisation. Skin punch biopsies were sliced into small pieces, covered with cover slips and cultured in a fibroblast medium. (A, B, C, D and E) Two different types of cells were observed between Days 4–7; the first were cells with a morphology consistent with that of keratinocytes (red arrows) while the second were cells with a morphology consistent with that of fibroblasts (black arrows). (F) Fibroblast cells became dominant in subsequent passages. (G) Beta thalassaemic mutations were confirmed in fibroblasts derived from beta thalassaemia patients using reverse hybridisation: 1) homozygous for IVSI-110 G>A, 2) normal control, 3 and 4) homozygous for IVII-1 G>A.

Emerging colonies were picked for further passaging. Colonies with a high nuclear-to-cytoplasmic ratio and compact edges were considered as possible iPSC clones. More colonies were obtained from the C19 than the C4 combination of episomal vectors (Table 3-1). TRA-160 immunostaining was used to discriminate between putative iPSCs and pseudo-iPSCs (Figure 3-2.B). A total of 11/138 and 19/91 of the screened clones were positive for TRA-160 in cells reprogrammed with the C4 and C19 combinations of vectors, respectively (Table 3-1). More iPSC clones were obtained when the three-vector combination (C19) was used, which is in agreement with Yu *et al.*[191]. The difference in proportion of iPSC clones generated by different vector combinations was statistically

significant (p=0.031, Fisher exact test). This indicate that the type of vector combination used is associated with a better rate of reprogramming to pluripotency.

Table 3-1: Number of iPSC clones generated from each patient using C4 and C19.

			R 1			**R2**				
			Plasmid combination 4 (C4)			Plasmid combination19 (C19)			C4	
		Sample type	No. cells transfected	No. clones picked	No. TR160+ clones	No. cells transfected	No. clones picked	No. TR160+ clones	No. TR160+ iPSCs	No. clones for further analysis
1	T1	BT IVSI-110	$1x10^6$	0	0	$1x10^6$	1	0		0
2	T2	BT IVSII-1	$1x10^6$	65	0	$1x10^6$	30	2		2(C19)
3	T3	BT IVSII-1	$1x10^6$	24	1	$6x10^5$	30	5	1	3(C19), 2(C4)
4	N4	Control	$1x10^6$	49	10	$6x10^5$	30	12		3(C19)
	Total			138	11		91	19		10

R1: transfection 1, R2: transfection 2

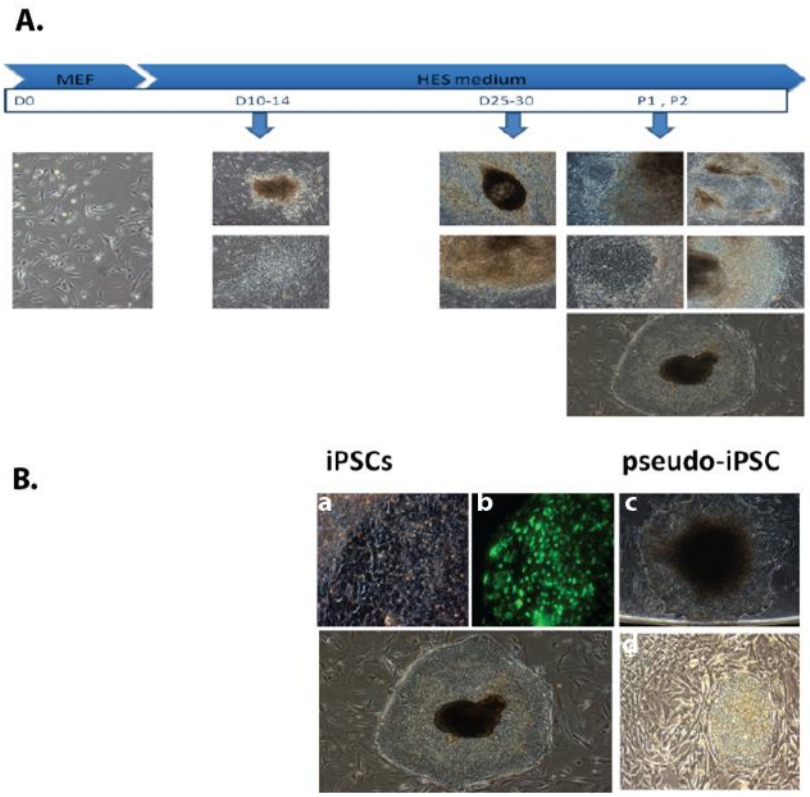

Figure 3-2: Timeline for iPSCs derivation from primary fibroblasts. (A) Colonies were observed on Day 10 after transfection and picked around Day 25–30. Rare colonies assumed an iPSC-like morphology before passaging, while the majority required passaging before they appeared to be characteristic of iPSCs. Most of the iPSCs generated during the course of the experiment for this study emerged between Passages 1 and 2. (B) TRA-160 staining was used to discriminate between putative iPSCs (b) and pseudo-iPSCs (c) and (d).

While these data show that the generation of footprint-free beta thalassaemia iPSCs from fibroblasts is possible, some patient fibroblast cultures were more refractory to reprogramming than others. This may be related to the age of the individual, the quality of the biopsy, genetic or epigenetic variability or may require further optimisation of the reprogramming conditions (e.g. increasing the amount of *SV40LT* and/or use *L-MYC*). For example fibroblasts from Patient T1, with the IVSI-110 mutation, repeatedly failed to reprogramme with either of the two plasmid combinations used (Table 3-1). Moreover, fibroblasts from Patient T2 (IVSII-1G>A) failed to reprogramme with the C4 combination; however, two clones were generated from this patient using the C19 combination. A total of seven putative iPSC clones were obtained from Patient T3 (IVSII-1G>A): two using the C4 combination and five from C19. Twelve and ten putative iPSC clones were generated from the control patient (N4) using the C19 and C4 combination, respectively. A maximum of 10 iPSC clones were selected from the normal line and all clones generated from beta thalassaemia patients fibroblast were expanded and cryopreserved.

Following the expansion of selected clones for seven to eight passages (Table 3-1), karyotype analysis revealed that both clones from Patient T2 displayed genetic abnormalities with a balanced reciprocal translocation involving the long arms of chromosomes 4 and 11 for one clone, and a translocation between chromosomes 13 and 15 for the other (Figure 3-3.A). One of the clones from the control subject, N4, showed monosomy for 8p and partial trisomy for a material of unknown origin (Figure 3-3.A). All karyotypically-abnormal clones were excluded from the study. Thus, further characterisation was performed on the five clones from beta thalassaemic patient T3 and three clones from the control subject, N4. The presence of the beta thalassaemia mutation in the T3-derived clones was verified by DNA sequencing confirming that iPSC clones were indeed derived from T3 beta thalassaemia cells (Figure 3-3.B).

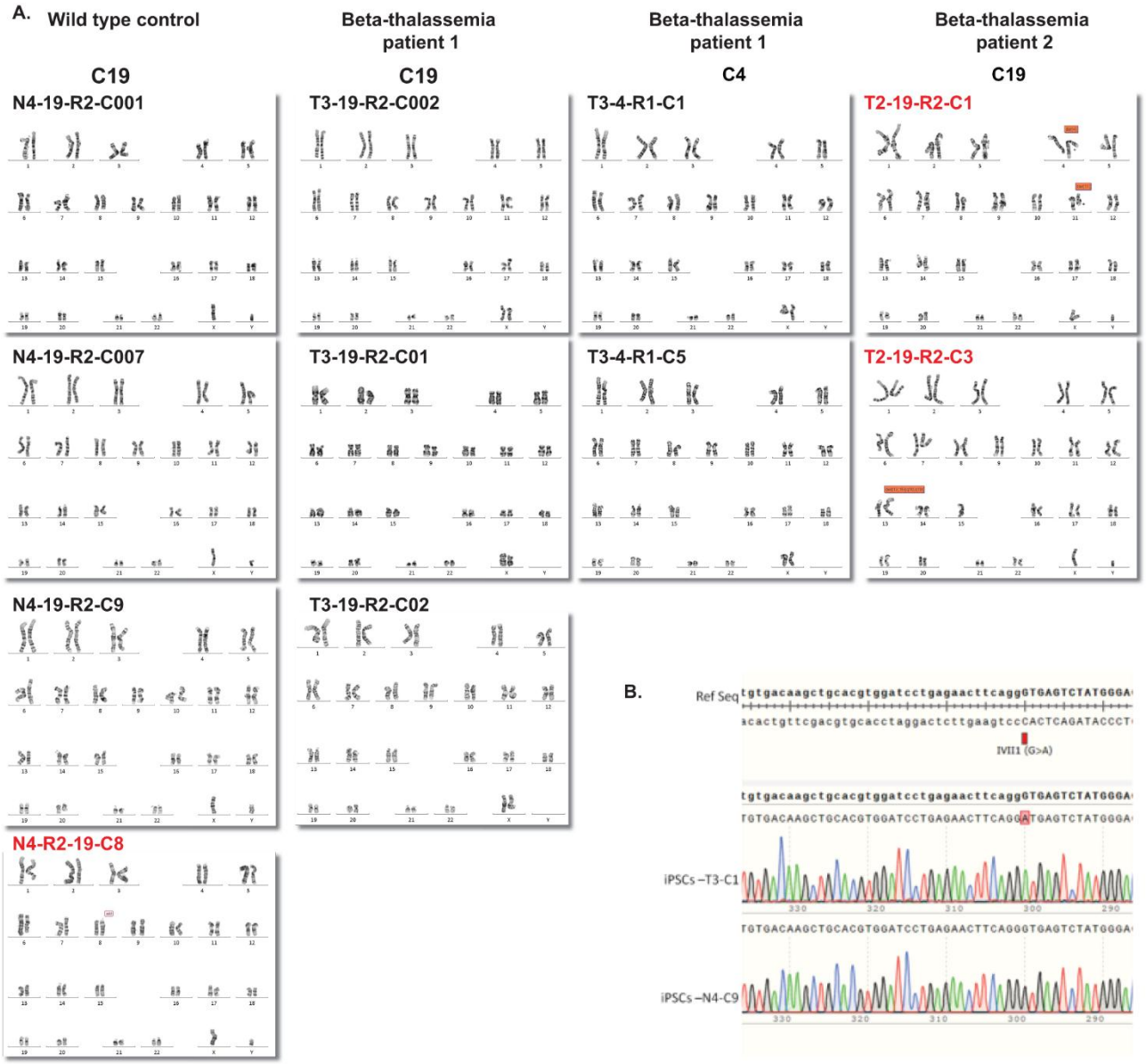

Figure 3-3: Characterisation of iPSCs: Karyotype/presence of the mutation. (A) Assessment of structural and numerical chromosomal abnormalities using standard G-banding karyotype analysis. (B) The presence of the beta thalassaemia mutation was verified in iPSC clones derived from beta thalassaemia patient T3 by DNA sequencing.

3.2.2 Assessment of the pluripotency of the iPSC clones

Both the T3 and N4 clones displayed robust expression of cell surface markers TRA-1-60 and TRA-1-81, as well as the pluripotency markers OCT4/POU5F1 and NANOG, as determined by immunocytochemistry (Figure 3-5). They also expressed endogenous *NANOG* and *OCT4* mRNAs at levels comparable to control H9 hESCs (Figure 3-4). Although a significant reduction in *SOX2* expression was observed in two iPSC clones (N4C9 and N4C007) compared to hESCs, the parallel reduction in *NANOG* and *OCT4* expression may indicate the observed decrease was caused by the

presence of differentiated cells. In fact, some clones were difficult to maintain at early passage numbers and improved at a later passage (p15).

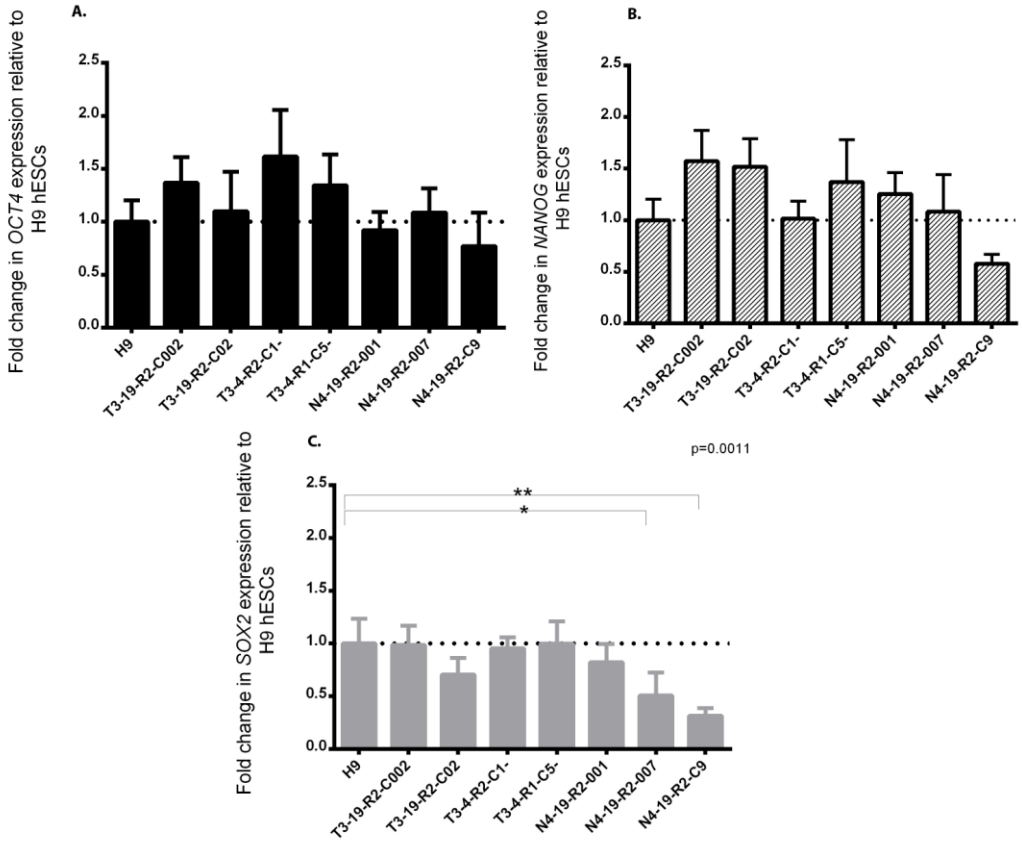

Figure 3-4: Expression of pluripotent marker genes in iPSCs. Fold change in expression of endogenous pluripotent cell markers A) *OCT4*, B) *NANOG*, C) *SOX2* genes for the indicated clones relative to H9 (normalised relative to *GAPDH*) by q-PCR. (Data represent mean±SEM, n=3). * $p<0.05$, ** $p<0.01$.

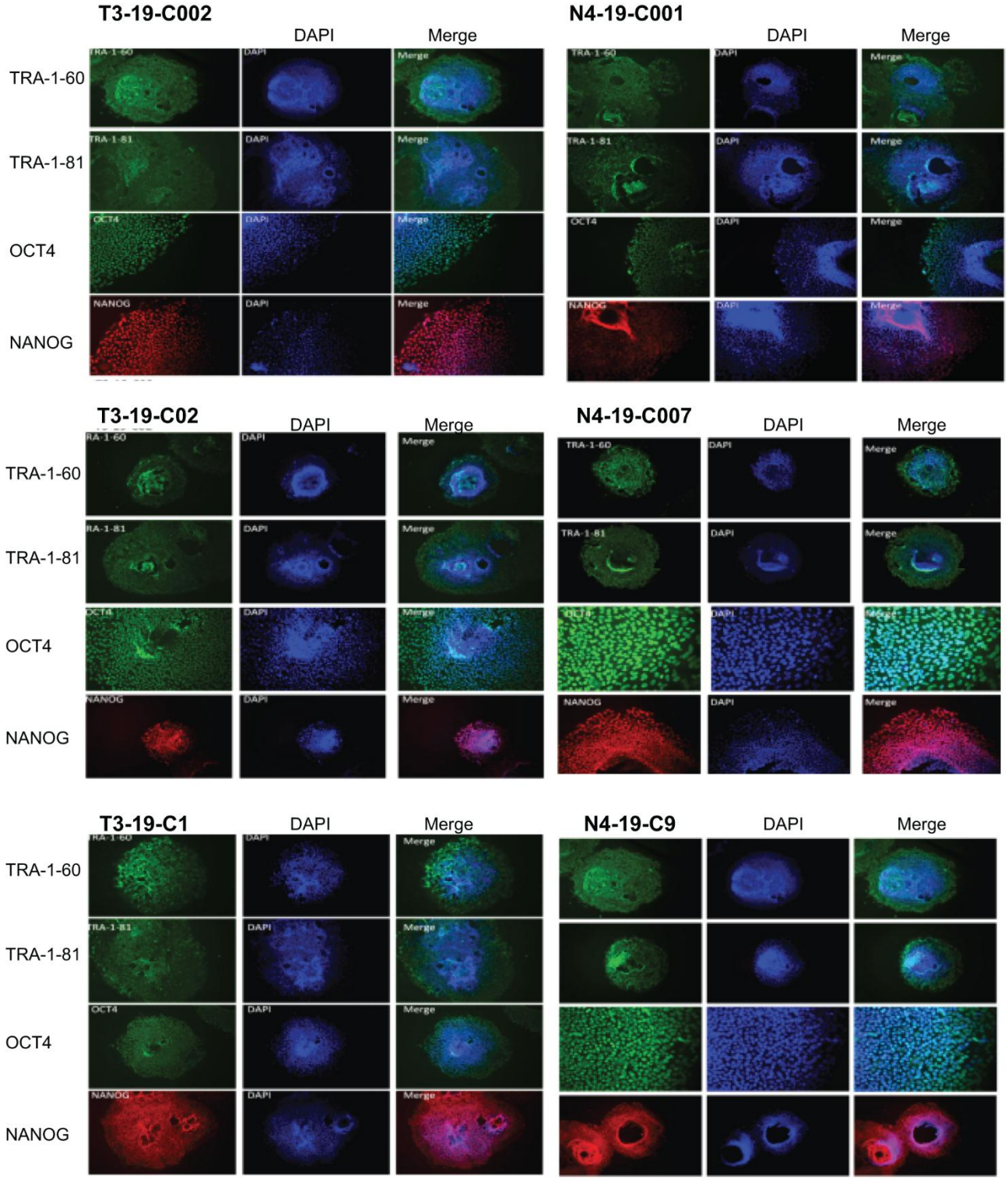

Figure 3-5: Immunocytochemical detection of the pluripotency markers OCT4, NANOG, TRA-1-81 and TRA-1-60 in control and beta thalassaemia iPSCs.

However, the robust expression of the pluripotency marker genes at the protein level (Figure 3-5) supports the notion that the previous observation most probably reflects the presence of differentiated cells at the time of collection. Next, spontaneous differentiation of iPSCs was assessed by transferring clumps of cells into a low-attachment cell culture platform in the absence of b-FGF in order to form EBs. EBs from all the tested clones expressed *GATA4*, *SOX17*, *SMA*, *Brachyury*, *PAX6* and *MAP2* at different levels (Day 23), indicating the pluripotency of the iPSC clones (Figure 3-6).

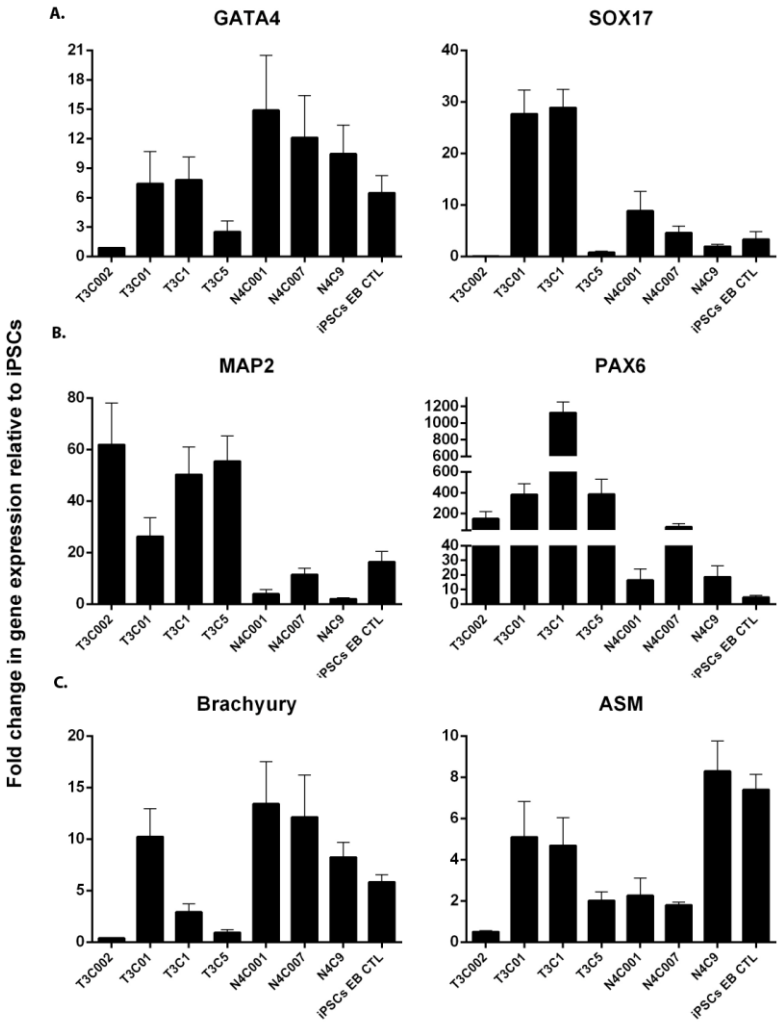

Figure 3-6: Expression of various lineage-specific genes in D23 EBs. Fold change in (A) endodermal (*GATA4* and *SOX17*), (B) ectodermal (*MAP2* and *PAX6*) and (C) mesodermal (*Brachyury* and *alpha smooth muscle actin*) markers during EB differentiation of clones from iPSCs lines (Data represent the mean±SEM, n=2).

Moreover, the expression of alpha smooth actin (SMA, mesoderm), alpha-fetoprotein (AFP, endoderm) and nestin (NESTIN, ectoderm) was detected using immunofluorescence in Day 23 EBs (Figure 3-7). AFP was not detected in one of the clones. However, pluripotency of all six clones was confirmed by their ability to form teratomas containing cell types representative of the three

germ layers following intramuscular injection into NOD-SCID mice (Figure 3-8), which indicates that the lack of the endodermal marker AFP was due to technical, rather than biological, reasons.

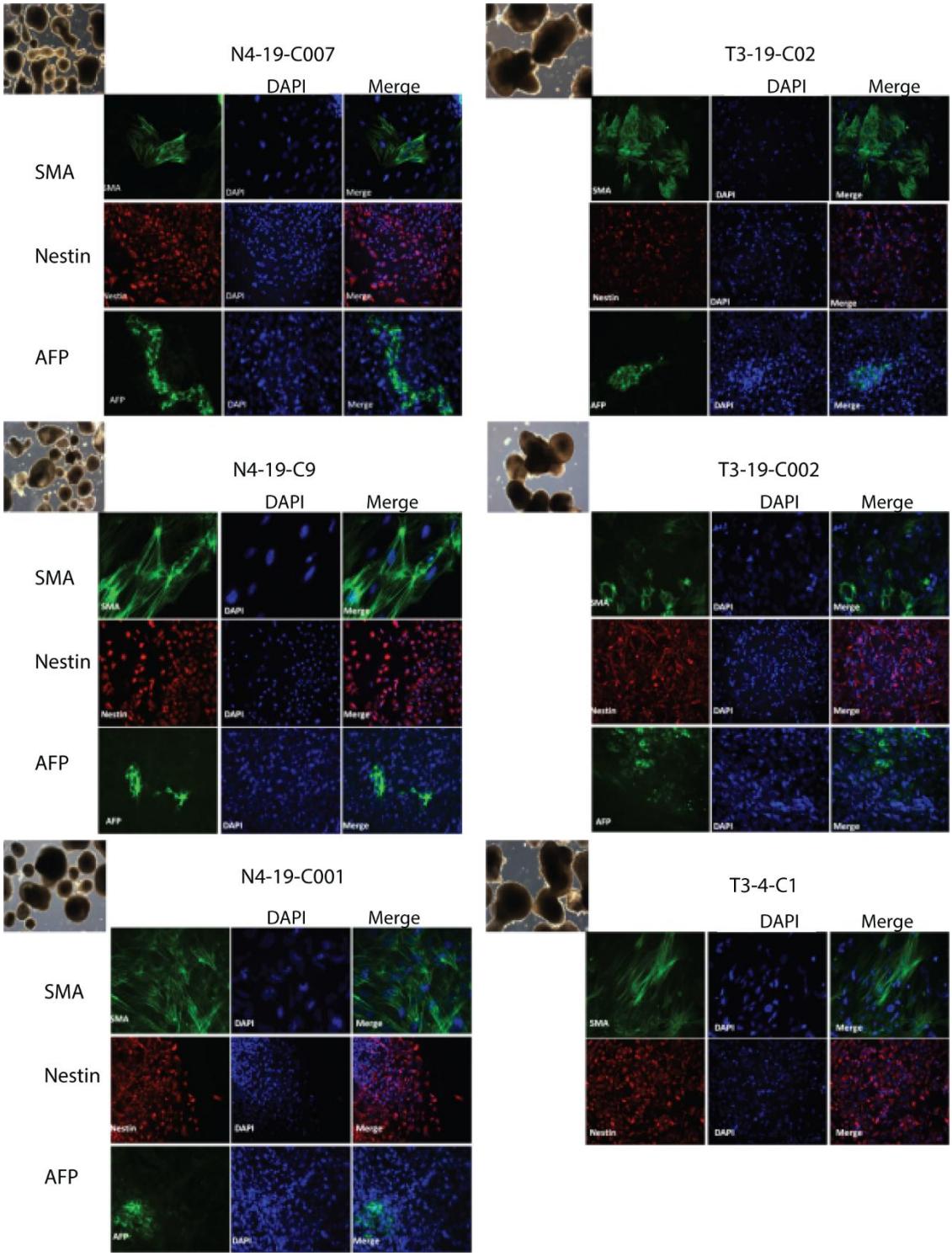

Figure 3-7: Immunostaining for lineage-specific markers in Day 23 EB after dissociation and plating on coverslips. Markers assessed were: alpha smooth muscle actin (mesoderm), alpha-fetoprotein (endoderm) and nestin (ectoderm) with DAPI nuclear staining for the indicated iPSC clones. (Alpha-fetoprotein was not detected in clone T3C1).

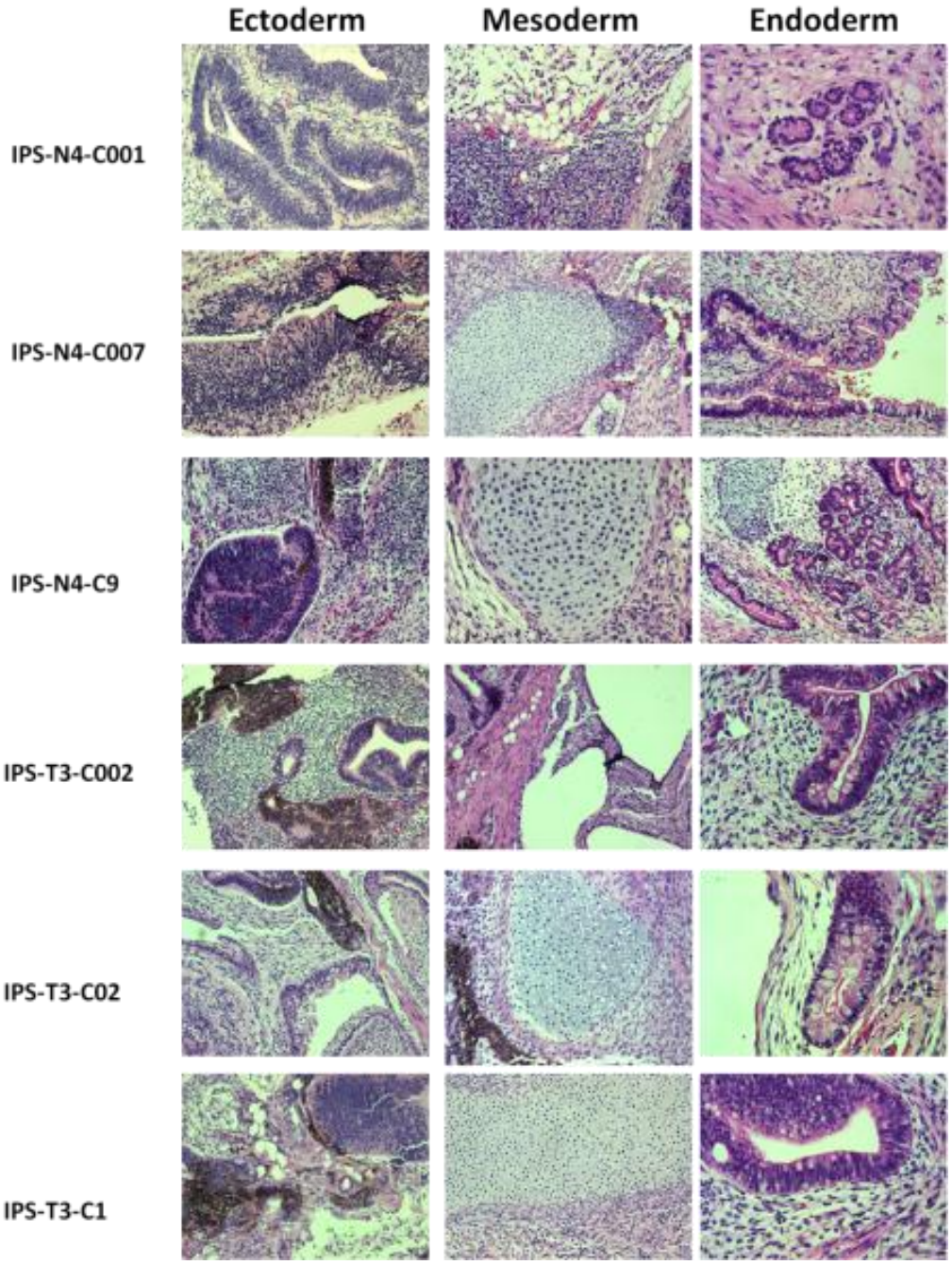

Figure 3-8: Teratomas demonstrating the trilineage competence of the iPSC lines. Haematoxylin and eosin histology images of derivatives of the three germ layers found in teratomas formed after injection of control and beta thalassaemia iPSCs into immunocompromised NOD-SCID mice.

3.2.3 Analysis of vector integration and persistence

To assess the absence of integration or persistence of the episomal reprogramming factors, PCR and RT-PCR analysis was performed on genomic DNA and RNA isolated from the iPSCs using transgene-specific primers. Episomal plasmid DNA could not be detected in five of the six iPSC lines, while transgene expression could not be detected in any of the clones (after excluding clone N4C007). The presence of amplicons for *EBNA*, *OCT4*, and *KLF4* in the absence of a positive band for *NANOG* and *LIN28* in one clone derived from the control fibroblasts indicates the OSTK plasmid was integrated in the genome (Figure 3-9.A, C007). The faint band for *EBNA* in the absence of the expression of reprograming factors in one clone may indicate a DNA contamination (Figure 3-9.C, C007). Although, most of the clones were integration-free, this finding highlights the importance of careful screening of the clones for the presence of a random integration. Of note, the presence of *GAPDH* amplicon in the sample of plasmid extracted from Day 15 transfected fibroblasts (positive control: PC1 and PC2) indicates contamination of the extracted samples with genomic DNA (Figure 3-9.A, and B).

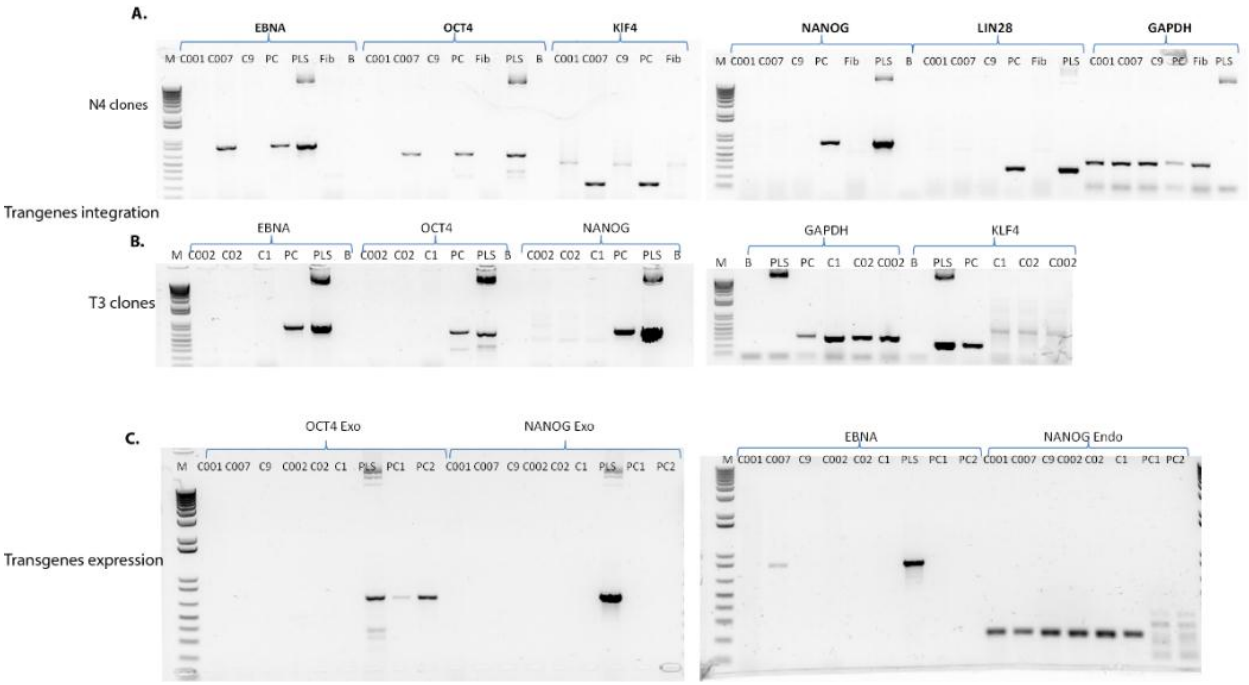

Figure 3-9: Presence and expression of transgenes as assayed by PCR analysis of genomic DNA and RNA isolated from the iPSCs using transgene-specific primers. Episomal DNA could not be detected in (A) 2/3 N4 iPSCs clones and (B) 3/3 T3 iPSCs clones, (C) while transgene expression was not detected in 5/6 clones. The presence of an amplicon in Lane 3C (C007) possibly indicates a low level of DNA contamination as no amplification was detected after treating the sample with DNase. (M: Marker, PC1: Plasmid DNA extracted from Day 15 transfected fibroblasts (T1), PC2: Plasmid DNA extracted from Day 15 transfected fibroblasts (T2), B: Blank control (no DNA), PLS: plasmid).

Thus, it is possible to generate *bona fide* high quality iPSCs from adult fibroblast samples from control and beta thalassaemia patients using episomes; however, there is a high attrition rate due to karyotypic abnormalities and some loss due to low-level transgene integration. Moreover, some patient samples are much more difficult to reprogramme than others, for reasons that are not immediately obvious.

3.3 *Conclusion*

This chapter describes the generation and full characterisation of footprint-free iPSCs from fibroblasts of individuals with and without beta thalassaemia using two different combinations of episomal vectors. More partially-reprogrammed iPSC clones emerged using C19 compared to C4. This difference can be attributed to variation in transgene arrangement that may impact their expression levels with subsequent consequences on the reprogramming processes [45, 192] as both vector combinations carry the same number and type of transgenes. Furthermore, even though episomal vector-based reprogramming is generally considered an integration-free process, careful screening of clones is needed to confirm the absence of transgenes in the genomic DNA, in agreement with previous reports [45, 46]. This is important since the transgenes that were silenced in iPSCs, as shown by RT-PCR, can be reactivated upon differentiation, which may limit the developmental potential of iPSCs and present further complications for future clinical applications in particular [193].

The nature of the vectors used to express the reprogramming factors is of critical importance for the generation of safe iPSCs. Despite the fact that reprogramming with episomal vectors is limited by a relatively low efficiency of reprogramming compared to other integrative methods, it only requires a single transfection followed by passive removal of the vector by continuous passaging. More importantly, this approach has been shown to generate iPSCs that are less immunogenic than those produced by retroviral delivery of the reprogramming factors after syngeneic transplantation of iPSC-derived cells in non-human primates [33]. It was also associated with a reduced incidence of teratoma formation by iPSC derivatives compared to clones produced by retroviral transduction [31]. This finding may be attributed to a greater presence of cells with a defective ability to differentiate (Type I phenotype) in retrovirally transduced cells as compared to those transfected with episomes. As such, the generation of adult fibroblast-derived footprint-free patient-specific iPSCs using episomes, as performed here, is another essential step towards a cell-based therapy for beta thalassaemia patients.

CHAPTER IV:

DOUBLE NICKASE-MEDIATED SEAMLESS REPAIR OF *HBB* MUTATIONS IN BETA THALASSAEMIA INDUCED PLURIPOTENT STEM CELLS

4 DOUBLE NICKASE-MEDIATED SEAMLESS REPAIR OF *HBB* MUTATIONS IN BETA THALASSAEMIA INDUCED PLURIPOTENT STEM CELLS

4.1 *Introduction*

Recent advances in genome engineering provide powerful gene-editing tools that enable site-specific DNA modification through the exploitation of the intracellular repair machinery [8]. Similarly, iPSCs constitute a powerful platform for providing a renewable source of cells with patient/disease-specific genetic backgrounds. Combining both tools and achieving site-specific editing in PSCs holds great potential for exploring the pathological processes associated with genetic disease, devising novel and effective therapeutic strategies and generating edited cells with potential therapeutic utility [8, 194]. However, achieving highly precise genome editing with engineered nucleases remains a challenge.

Type II CRISPR/Cas9 is one of the most prevalent nuclease systems that is widely adopted for genome engineering. It is characterised by the ease of its design, its efficiency and the possibility of repurposing this system to carry additional molecular functions [156]. The system is dependent on a modular short guide sequence to direct the nuclease domain to cleave its DNA target site provided that a PAM motif sequence site is located downstream of the protospacer sequence [59]. As a result of the induced DNA break, the DNA repair mechanisms are stimulated and either the error-prone NHEJ, or the HDR pathways predominate, depending on the gene editing strategy employed. In fact, in the presence of an ectopic donor plasmid, the repair can be guided to tailor the DNA modification through HDR [195]. Subsequently, insertions, deletions, and point mutations can be formed at the targeted site with various consequences including, gene disruption, gene correction, transgene insertion or a combination of simultaneous modifications. However, unwanted on-target [145, 196] and off-target induced mutagenesis [58, 65, 74, 84] are frequently associated with Cas9 activity. On-target mutagenesis is caused by the iterative cycles of Cas9 activity after the introduction of the desired genetic modification (i.e., unwanted specific activity), whereas off-target mutations are mostly caused by the tolerance of Cas9 to sequence mismatches which, results in non-specific mutagenesis (i.e. non-specific activity). The target sequence properties, the type of nuclease variants, and targeting strategy, are all factors that influence the repair pathway activated (i.e. accurate or mutagenic pathways).

To limit off-target activity, an approach that biases the resolution of DNA breaks through high fidelity repair pathways has been suggested [95, 96], that is induced by the nickase variant of Cas9 and is supposed to be quickly and efficiently repaired through either the faithful BER mechanism [97] or error-free HDR [58, 63, 118, 119], while reducing the frequency of the error-prone NHEJ [101, 119]. However, nickase-induced HR has also been reported to display inefficient activity that

can be sequence- and locus-dependent [117, 145]. Further improvement of nickase efficiency was achieved by adopting a 'paired nickase' strategy. This strategy has been suggested to favour a more effective on-target modification (compared to a single nickase), while reducing off-target effects [95, 96]. This was based on the notion that paired nicks induce a composite DSB at the targeted site, while only a SSB is induced at off-target sites. However, HDR and NHEJ compete and cooperate to repair the composite DSB and thus are thought to lead to undesired on-target modifications following gene correction [118, 197]. To address this concern, a donor strategy design was suggested that converts a DSBs to a SSBs at the targeted allele following HR. This was achieved by designing a donor template that inserted a selection cassette at either a one of the guide-binding sites (inactivating one guide) or between the guide-binding sites of a pair (both guides-binding sites are intact but the DNA nicks are not closely spaced) [145, 198]. However, for this strategy to be effective, the nicks closely positioned in the targeted allele must be resolved by a high fidelity repair mechanism.

The data on the mutagenicity of monomeric nickases are inconsistent at present. Moreover, the majority of the studies on nickase activity to date have been carried out in transformed, or immortalised, cell lines and limited data are available on the efficiency and safety of this strategy in the context of gene correction in human pluripotent stem cells. The prevalence of on-target indel levels (NHEJ) induced by a single nickase was reported to range between undetectable [96], measurable (low activity) in HEK 293 cells [58, 61] and U2OS cells [98] or considerable activity in HEK 293 cells [71]. Although these studies indicate that the type of repair can be sequence- and locus-dependent, they also indicate that SSBs can be repaired through an error-prone pathway at some guide-binding sites, which is clearly of concern when using this method for precise genome editing.

Different cell types can use different pathways for the repair of DNA lesions resulting in different outcomes. Generally, iPSCs exhibit a low off-target mutation load with Cas9WT as assessed by whole genome sequencing [88, 89] or screening of the top predicted off-target sites [35, 36, 77, 87]. However, this was disputed when researchers targeted Intron 2 of the *HBB* gene [138]. Similar inconsistencies were observed with monomeric Cas9n (single guide), which showed undetectable mutagenesis rates in bulk iPSC populations [58], whereas a high variability of monomeric nickase activity at the targeted allele was evident in iPSCs obtained after positive selection but not in bulk cell populations [145]. Interestingly, despite the high activity of nickase observed at some loci, these authors showed that abolishing the binding site for one guide, or separating the DNA nick sites by a selection cassette in the context of a double nickase strategy can considerably reduce on-target mutagenesis in iPSCs.

Assessing strategies based on a DN approach with respect to gene correction of a disease-causing mutation in human iPSCs is particularly relevant for a gene like *HBB,* which has been reported to have a considerable propensity for off-target activity for some sgRNA sequences [138, 139]. In this study, it was hypothesised that a Cas9-DN approach for *HBB* gene correction would benefit from strategies that: (1) convert DSBs to nicks at the target allele as this limits unwanted mutagenicity; (2) generate bi-allelic targeting, considering that the untargeted site will be vulnerable to NHEJ-mediated mutagenic events; and (3) approaches that allow for both positive and negative selection mode, so as to facilitate the identification of targeted clones and permit subsequent seamless excision with the aim of preserving genomic integrity. Therefore, the present study sought to assess the applicability of a DN approach to achieve biallelic seamless gene correction of a disease-causing mutation in iPSCs derived from patient with beta thalassaemia major (homozygous), and evaluate and compare on-target mutagenesis in iPSCs using two different design strategies for corrective gene targeting. In both strategies, the double nickase activity on the targetted allele was eliminated following HR by designing a donor template that either introduced a SNP in the PAM sequence at the binding site for one of the guides (sgRNA1) of the first pair set (sgRNA1/3), or by inserting a large selection cassette flanked by the *piggyBac* transposon inverted repeat (>3 kb) at the binding site for another guide (sgRNA4) in the second pair (sgRNA45).

The results revealed that biallelic seamless gene correction of a beta thalassaemia-causing mutation is possible, and that the vast majority of the monoallelically targeted clones (after excluding random insertions) possess an indel at the guide binding site of the untargeted allele (original allele), indicating a high level of on-target activity that can be attributed to either double, or in some cases a single nickase activity. Upon further investigation of the effect of this approach on reducing on-target activity at the targeted allele, it was found that there was no indel associated with the targeted allele obtained through modification of the PAM sequence, and destroying the binding site for one guide/separating the guides in the other pair did abolish the DN activity but was unable to fully eliminate DNA modification as revealed by the presence of an indel at the binding site of the second guide of the pair in 5 out of 26 of the screened clones, strongly suggesting existence of an unexpected nickase-induced activity associated with one guide.

4.2 *Results and discussion*

The following sections will present and report on:

1) The strategy used to design various gRNAs to target exon/intronII of the *HBB* gene, the validation of this approach in HEK293FT cells using Cas9WT, the assessment of the functionality of the first guide pair (sgRNA1-Cas9WT and sgRNA3-Cas9WT), and the evaluation of the paired nickase activity of five multiple pairs ($sgRNA_{A/B}$-Cas9n) in HEK293FT cells.
2) Optimisation of the iPSC transfection conditions, using sgRNA1-Cas9WT-EGFP, to maximise the transfection efficiency (via GFP-positive cells quantifications) relative to cell viability.
3) Targeting the *HBB* gene with a *piggyBac*-based donor construct and Cas9DN using two different strategies that convert DSBs into SSBs upon targeting: a sgRNA1/3-Cas9n+ donor (inactivation of PAM motif for sgRNA1) and the sgRNA4/5-Cas9DN+ donor (separation of the guide pair by inserting the selection cassette).
4) Screening for HR events and on-target mutagenesis after positive selection with puromycin.
5) Validation of the correction of the disease-causing mutation by sequencing.
6) Assessment of off-target activity at five top predicted off-target sites per guide.
7) Assessment of the restoration of the wild type allele after seamless excision of the selection cassette.
8) Evaluation of the pluripotent state and karyotype analysis of the corrected cells.

4.2.1 Gene targeting of the *HBB* locus with Cas9 or Cas9n in HEK293FT

4.2.1.1 *Design of guide RNA pairs for genome editing using a double nickase approach*

The efficiency of DNA modifications (through NHEJ or HDR) in response to composite DNA breaks induced by paired nickase is affected by the sequence context, the offset distance between pairs of guides used to target the intended region, and the distance of the nicks relative to the intended modification [96, 145, 146]. The intended DNA modifications in the current study involve a corrective substitution of the disease-causing mutation and an insertion of an excisable dual selection cassette using the *piggyBac* transposon-based approach described in Yusa *et al.* [182]. Therefore, multiple pairs of sgRNAs were designed to screen for efficient guide pair capable of introducing DNA breaks at different distances relative to either the insertion or the targeted mutation. Some of the pairs were designed to introduce a DNA nick close to the *HBB* mutation site whereas other pairs introduced the break upstream to the TTAA insertion site (for *piggyBac*). The offset distance for different pairs varied between -7 to 51 bp (Figure 4-1.B**)**. Vectors with dual

expression of sgRNA and Cas9 nuclease (or Cas9-D10A nickase) were used because this ensures concomitant expression of both components in transfected cells.

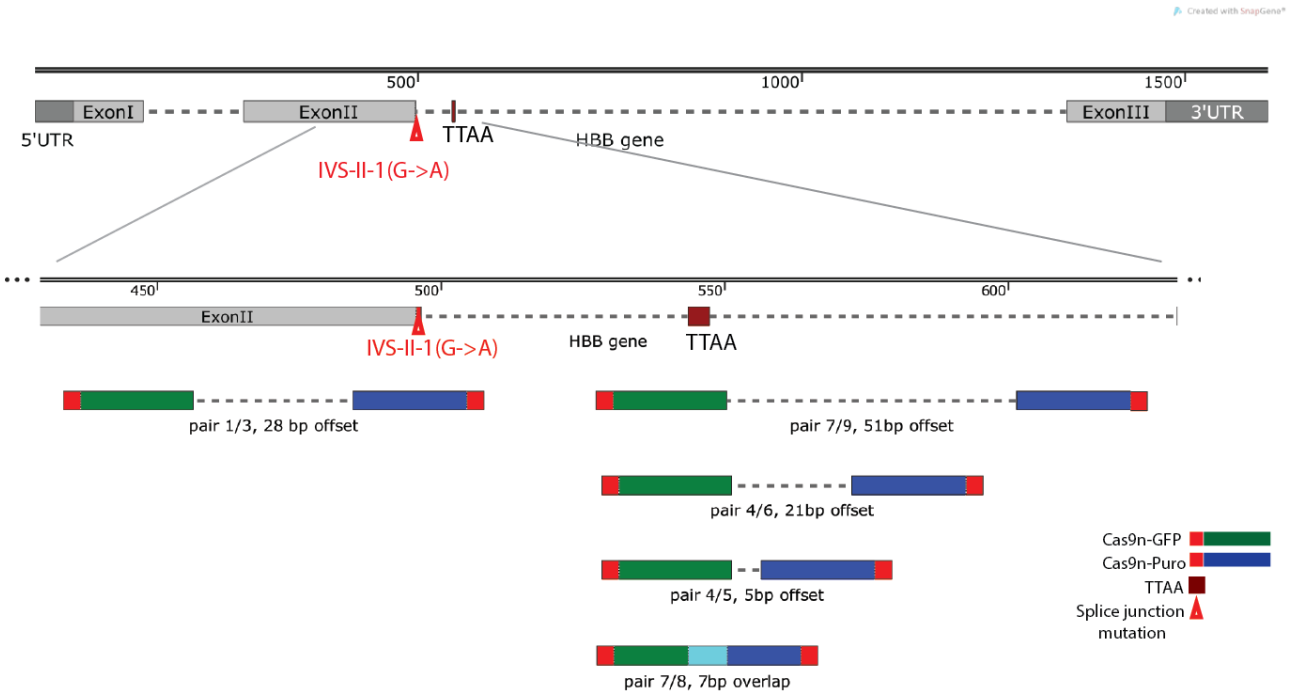

Figure 4-1: Schematic illustrating the HBB gene and the location of multiple pairs of sgRNA. Guides with a different offset (dashed line) were designed to either target the region around the mutation site (Δ) or the TTAA site (dark red box). The first sgRNA of the pair was cloned in an EGFP expression vector (green), while the second was cloned in a puromycin-expressing vector (blue). The PAM sequence is indicated by the red box.

4.2.1.2 *Assessment of the efficiency of sgRNA-Cas9(n) targeted genome modification in HEK293FT cells*

To assess if the system could generate indels at the *HBB* locus, the efficiency of inducing indels at the gRNA binding sites was initially evaluated for one pair of guides using a Cas9WT nuclease-expressing vector (Figure 4-2.C). Guides for the first pair (gRNA1 and gRNA3) were cloned into a vector co-expressing a human codon-optimised Cas9WT nuclease and puromycin (PX459). One guide (gRNA1) was also cloned into a sgRNA1-Cas9WT-EGFP (PX458) expressing vector to assess the transfection efficiency [177]. Vectors were independently transfected into HEK 293FT cells using Lipofectamine LTX. Cells transfected with the PX459 vector were subjected to puromycin selection for 48 hrs. Transfection was determined as being efficient as >80% of the cells were GFP-positive as identified by GFP-positive cells using fluorescent microscopy (Figure 4-2.A). DNA was extracted 72 hours post-transfection to assess the rate of targeted DNA modification. The indels resulting from NHEJ were analyzed with the Surveyor assay using PCR amplicons amplified with SV1 primers (Table 2-2). The size of the fragment generated by the Surveyor nuclease

corresponds to an indel at the expected Cas9WT cutting site (Figure 4-2.E). The efficiency of indel frequency (as quantified using Imagelab software, BioRad) induction by NHEJ was 28.78 per cent in the absence of selection. However, this value increased two-fold with drug selection, reaching 51.97 and 58.3 per cent for sgRNA1 and sgRNA3, respectively (Figure 4-2.C).

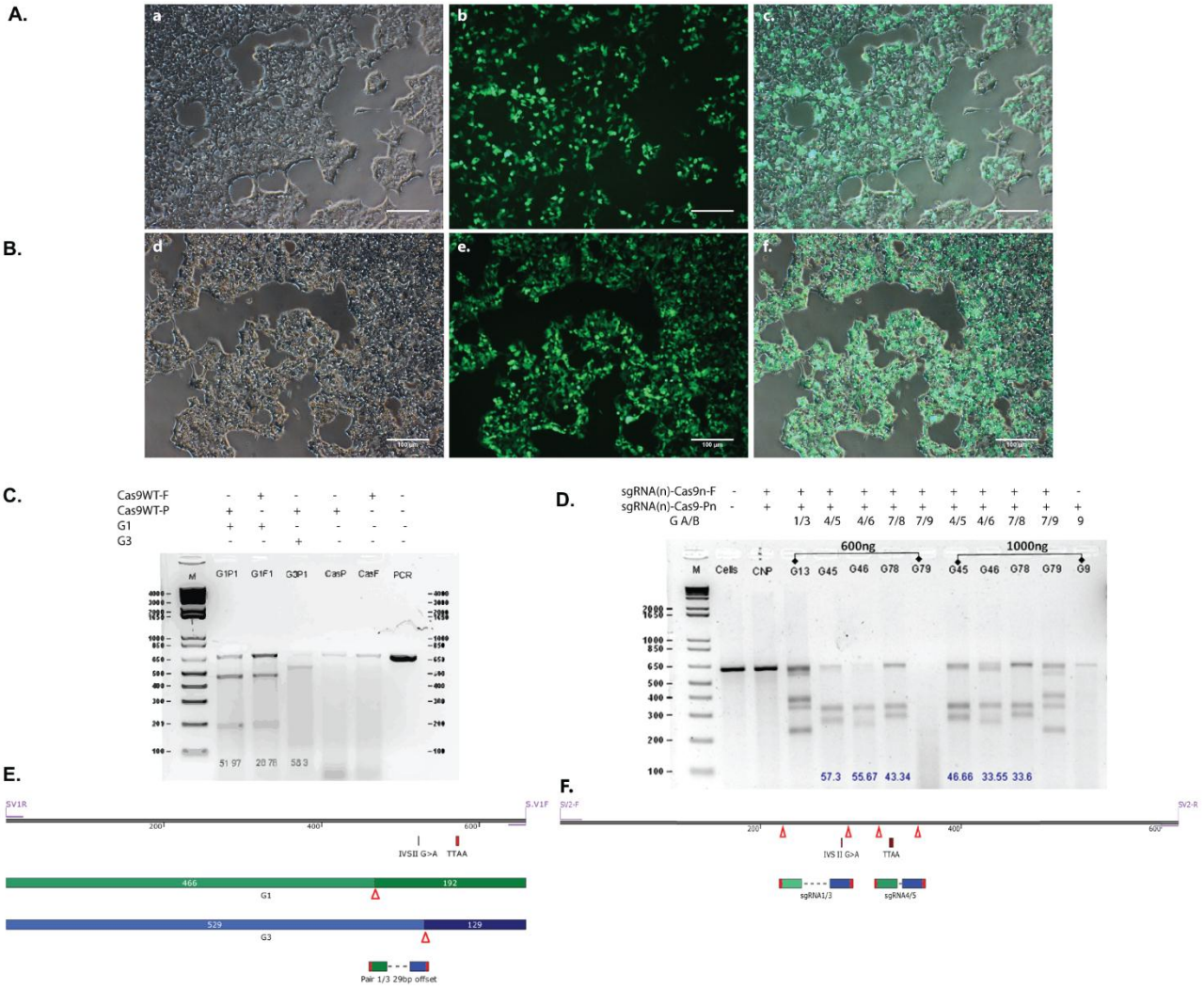

Figure 4-2: Functional analysis of guide RNAs using Cas9 or Cas9DN systems in HEK293 cells. (A) Transfection of HEK293FT cells with sgRNA-Cas9-EGFP-expressing vector (B) or co-transfection of both sgRNA-Cas9n-EGFP and sgRNA-Cas9n-puromycin-expressing vectors resulted in efficient GFP expression as displayed by fluorescent microscopy (b and e). Phase contrast image is shown in a and d and merge image in c and f. The transfection resulted in efficient on-target modification as demonstrated by the Surveyor assay for (C) Cas9WT (D) and Cas9 double nickase systems for the indicated guide pair. (E) Schematic illustrating the location of the Cas9 cleavage sites of guide1 and guide3 (red triangle) for primer SV1 and the predicted size from the Surveyor nuclease (F) and the predicted cleavage sites for both sgRNA1/3 and sgRNA4/5 for primer SV2. Primer pairs SV1 and SV2 were used in C and D, respectively.

This observation possibly indicates that Cas9WT did not result in mutagenesis in nearly 40 per cent of the transfected cells (puromycin-resistant cells mostly expressing GFP).

Different sgRNAs can have different activities [146]. This is of relevance when a composite DNA break is needed. Therefore, for the paired nickase approach to be effective, both guides forming the pair should ideally have comparable activity to efficiently form a staggered composite double DNA break. All five pairs were next assessed for their ability to induce indels at the targeted site (Figure 4-2.D). Two different expression vectors were used to clone guides for each pair; PX461 (EGFP) and PX462 (puromycin). This strategy facilitated monitoring of the transfection efficiency while allowing for drug selection. The effect of paired Cas9-nickase dosage on the indel efficiency was also explored using either 600 ng or 1000 ng of total plasmid. Transfected cells were subjected to drug selection (for two days) starting at 24 hours post-transfection. The majority of puromycin-resistant cells were positive for GFP, as observed under a fluorescence microscope (24 hours after selection) (Figure 4-2.B). This observation indicates that efficient co-transfection of the vectors was achieved. The induced indel efficiencies obtained for vectors expressing sgRNA45, sgRNA46, sgRNA78 were 57.3, 55.67 and 43.34 per cent for 600 ng of pDNA and 46.66, 33.55 and 33.6 per cent for 1 μg of pDNA, respectively. These values indicate that increasing the plasmid did not increase the efficiency. It is likely that the decrease in efficiency observed at 1 μg pDNA reflects a loss of transfected cells due to either plasmid load-induced toxicity or suboptimal transfection conditions. These results may also have been affected by the low sensitivity, low signal-to-noise ratio [199] and accuracy of the Surveyor assay. Moreover, differences in sample loading in different lanes or experimental variability associated with transfection condition can also bias the results. The cutting efficiency was not normalised to transfection efficiency as the transfection efficiency was consistently high in HEK293FT as judged by fluorescence microscopy. Moreover, the plasmid dosage effect was likely to differ in iPSCs, as they are a different cell type and require different transfection conditions (lipofection/nucleofection).

Upon further analysis, multiple bands were observed for sgRNA13 and sgRNA79. It is highly unlikely that this pattern would be the result of 5′ exonuclease activity of the Surveyor nuclease [200] as the band pattern (for the additional band) seems to be proportional to the offset distance between the pair of guides (Figure 4-3.B). In fact, diffuse PCR amplicons were observed for sgRNAs 46, 13 and 79 with an offset of 21, 28, and 51, respectively, and this was not observed for the offset of -7 and +5 bp (Figure 4-3.A). At first glance this band pattern may indicate a subpopulation of cells with indels resulting from the repair at a single nick. However, this hypothesis is highly unlikely as the bands generated with the single nickase are barely detectable above the background levels (Figure 4-2.D**,** G9). A similar observation was reported in another

study using paired nicks with offsets greater than 100 bp [61]. They reasoned that the DSB occurred in a region between the nicks during the process of repair which was resolved through a head-to-head strand replacement resulting in deletion [61]. Interestingly, in that study, a parallel band pattern was obtained with paired Cas9WT (compared to Cas9DN) which indicated the occurrence of a large deletion of the intervening sequence (confirmed by sequencing). Contaminants or spontaneous reversion of the D10A mutation in the RuvC I-Cas9 domain was ruled out after sequencing the Cas9n vectors in the current study.

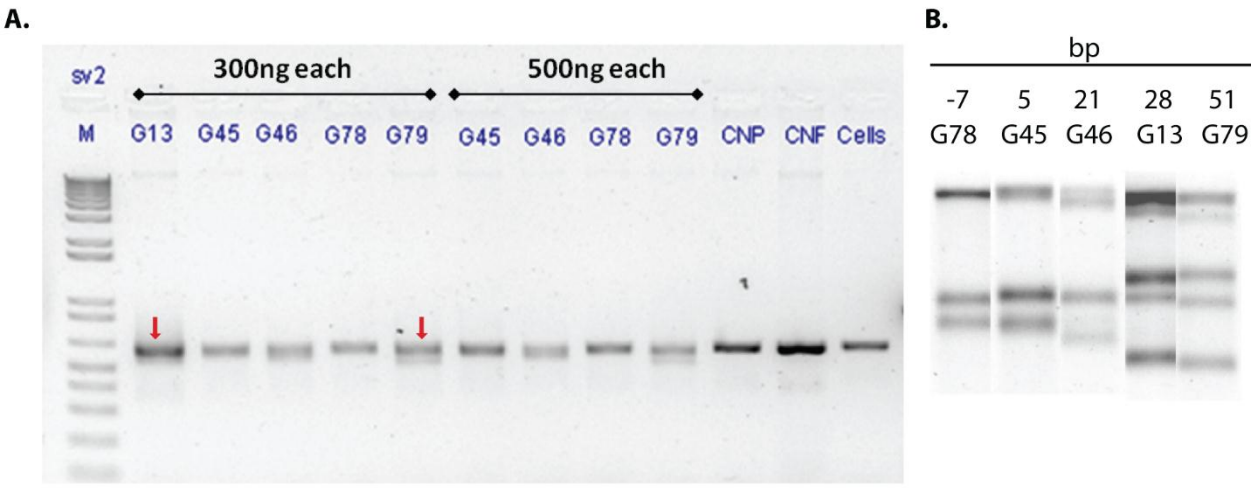

Figure 4-3: Band pattern for PCR amplicons, generated using SV2 from transfected HEK293 cells (Cas9DN), and their Surveyor nuclease treated products. The band pattern possibly indicates a large deletion that seems to be proportional to the offset distance between the pair of guides as indicated by: (A) The diffuse band for PCR products for offset distances of 28 and 51 bp (red arrows), (B) and the products after digestion with Surveyor nuclease for the indicated guide pairs. The gel in B was cropped and rearranged from (Figure 4-2.D).

Moreover, repair through the formation of a DSB intermediates was recently demonstrated for a DN strategy with an offset of less than 100 bp [118]. Thus, it was concluded that the overhang length may be a factor that possibly dictates a type of utilised repair mechanism at the composite break sites. The band pattern suggests the presence of two defined processes that resolve the composite break.

4.2.2 Optimisation of iPSC transfection conditions

4.2.2.1 *Liposomal-based transfection of iPSCs*

To deliver the sgRNA-Cas9WT-EGFP plasmid liposomal delivery approaches that do not require cell suspension for transfection, were explored using commercially available lipid-based transfection reagents. The reagents tested included Lipofectamine 3000 (Life Technologies),

Fugene HD (Promega), TransIT-LT1, and TransIT-2020 (Mirus Bio's). A low survival and very low transfection efficiency was observed upon transfecting iPSCs with Lipofectamine 3000 despite the fact that adherent iPSCs (single cell format) were used as opposed to cell suspensions (in nucleofection). Next, transfecting colonies as opposed to single cells was tested in an effort to improve the transfection efficiency. Transfection of colonies was carried out as per the manufacturer's instructions using two different charge ratios (1.5X and 3X) of Lipofectamine3000 to pDNA and 1µg of pDNA in 12 well plates.

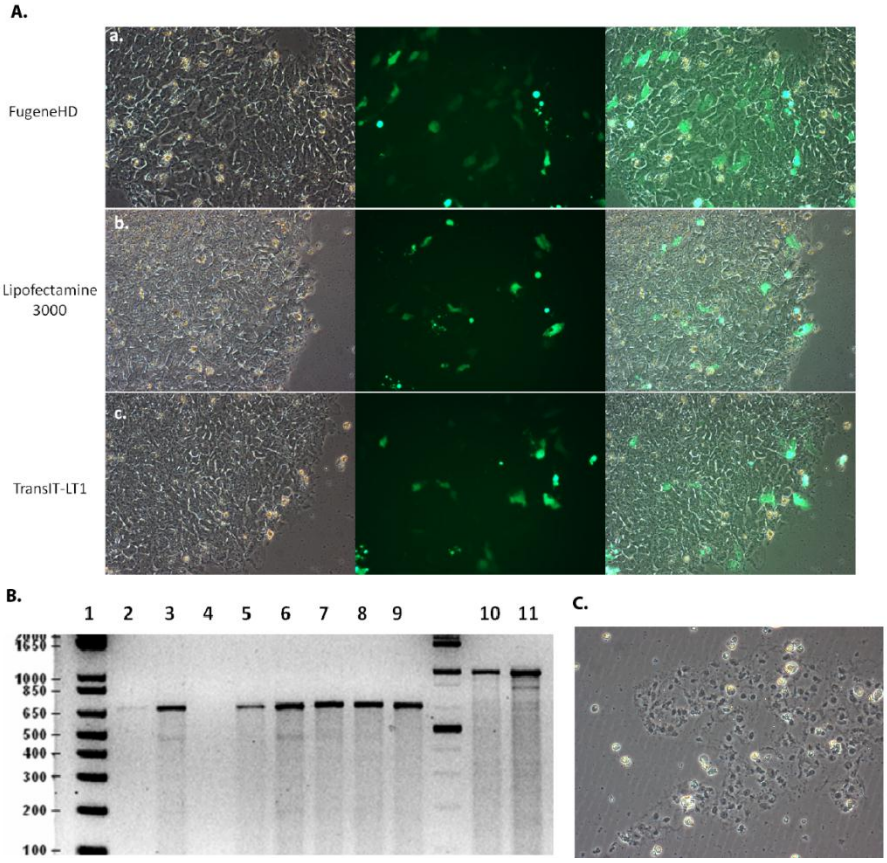

Figure 4-4: Liposomal-based transfection of iPSC clones with the sgRNA1-Cas9WT-EGFP-expressing vector. (A) Transfection efficiency for clones with the indicated reagent. (B) The calculated efficiency for on-target modification as assessed with Surveyor assays was 4.3 per cent (Lane 3), 7.3 per cent (Lane 6) using SV1 primers, and 6.1 per cent (Lane 11) using SV3 for TransIT®LT1, Lipofectamine 3000 (3X), and Fugene HD, respectively. (C) A toxic effect was observed on Day 5 for some clones. IPSC controls: Lanes 9 and 10.

Notably, fluorescence was visible between 24-72 hours post-transfection, and decreased considerably from Day 4 onwards. This finding indicates a short, transient expression of EGFP. After trialling all liposomal transfection reagents, it was found that transfection Fugene HD yielded the highest number of GFP-positive cells (as observed under a fluorescence microscopy) (Figure 4-4.A). Fluorescence-activated cell sorting (FACS) was not used to estimate the percentage of fluorescent cells, as a measurement of the efficiency of the induced indel at the targeted site can be

a more relevant indicator for downstream applications than estimating transfection efficiency by FACS. Thus, Surveyor nuclease assays was carried out to compare the transfection efficiency of the various reagents. Estimated induced indel efficiencies of 7.3, 4.3 (using SV1 primers), and 6.1 per cent (using SV3) were obtained for Lipofectamine 3000 (3X), *Trans*IT®LT1 and Fugene HD, respectively, under the reported experimental conditions and transfection with sgRNA1-Cas9WT-EGFP (Figure 4-4.B). The discrepancy between the observed fluorescence and induced indel efficiency for Fugene HD and for Lipofectamine 3000 may reflect the loss of transfected cells, as cell toxicity was observed in a subset of transfected iPSC clones (Figure 4-4.C). It may also reflect the high signal-to-noise ratio that influences the estimation of band intensity. No bands were observed for Lipofectamine 3000 (1.5X) and TransIT-2020, possibly indicating that the activity was below the detection limit of the assay. It was concluded that both Fugene and Lipofectamine 3000 can be used as transfection reagents for the delivery of the sgRNA-Cas9 plasmids. Further optimisation of experimental conditions regarding the amount of plasmid used and the charge ratio relative to cell density will be necessary to improve the observed indel efficiencies.

4.2.2.2 *Optimisation of nucleofection conditions*

For CRISPR-Cas9-mediated gene correction, induced PSCs that had been adapted to single cell feeder-free culture conditions were used. Despite trialling different programmes that included CB150 and CA137 (common stem cell-specific programmes) to nucleofect $5x10^5$–$1x10^6$ cells with 1 μg sgRNA1-Cas9WT-EGFP vector (P3 buffer, 4D Nucleofector), the iPSCs continued to exhibit low viability after transfection (less than 25 per cent viable cells), as estimated by trypan blue exclusion, when seeded under feeder-free conditions. This was coupled with a very low transfection efficiency, as judged by the number of GFP-positive cells. Similar findings were obtained using a different buffer (Ingenio Electroporation buffer) with two different transfection programmes, B016 or A023 (Nucleofector II, Amaxa) [201]. However, seeding the cells on MEFs after transfection yielded promising results. In fact, extracellular matrix (ECM) and irradiated MEFs were tried for plating cells subjected to nucleofection with sgRNA1-Cas9WT-EGFP. Cell survival improved with an increased plating density for cells seeded on ECM (Figure 4-5.B). However, a significant improvement in cell viability was observed when a layer of irradiated MEF was used to plate cells on after transfection, as judged by the number GFP-positive cells (Figure 4-5.A) and by the number and the morphology of clones that formed after seven days compared to ECM alone, even at a low plating density. Therefore, irradiated MEFs were used for plating transfected iPSCs. Further optimisation attempts were made using different cell density to total pDNA ratios, nucleofection programmes and seeding densities after nucleofection, as follows:

a) Programme for nucleofection: Cells ($5x10^5$) were subjected to six different nucleofector programmes that varied in strength and centred on CB150 (suggested by the manufacturer's instructions; weak CA137<CA-150<CB-141<CB-150<CD-150<DE-150<DE-148 strong) in 16-well nucleovette strip format with 1 μg pDNA. Transfected cells were plated on irradiated MEF plates (12-well) in the presence of ROCK inhibitors. Programme CB150 had a higher count of GFP-positive cells as observed by fluorescence microscopy (Figure 4-5.A).

b) Cell density to pDNA ratio: The CB150 programme was used to transfect 1–$2x10^6$ cells with a range of pDNA concentrations (1–4 μg). Although transfecting $2x10^6$ cells with 2 μg of pDNA produced the highest number of GFP-positive cells, the resulting indel efficiency (as assessed by Surveyor assays) was not significantly different compared to transfection with 4 μg/$2x10^6$ cells, which produced lower numbers of fluorescent cells. This may indicate that the system is tolerant to a range of conditions and that GFP expression does not necessarily correlate with indel efficiency. The modest difference in indel efficiency did not result from experimental or measurement artefacts, as it was consistent between two different PCR amplicons generated with two different set of primers (SV1 and SV3) from the same sample DNA. The observed indel efficiencies using 2 μg and 4 μg of pDNA were 10.6 and 9 per cent (with SV1 primers) and 7.07 and 5.77 per cent (with SV3 primers), respectively (Figure 4-5.C). The difference in the absolute values of indel activity between both primer sets from the same sample indicates that this assay can be affected by many variables, including the background of cleavage, especially at a low rate of mutation. Based on results from the observed GFP-positive cells and indel efficiency, 2 μg of plasmid was used to transfect $2x10^6$ cells. Overall, the indel efficiency achieved in HEK293FT cells was 2.8-fold higher than that observed in iPSCs for the same sgRNA-Cas9 vector.

Having optimised the parameters for iPSCs nucleofection, the next step was to assess the indel efficiency induced by paired nickase using guide pair 1 and 3 with two set of primers (SV1 and SV2). Although a band of the expected size was detected, the signal-to-noise ratio was high (Figure 4-5.D), preventing the estimation of the cutting efficiency.

These data suggest that the guides designed were functional in both the Cas9WT and Cas9 paired nickase strategies, as judged by the presence of indels at the targeted site of the *HBB* gene in iPSCs. It appears that under these experimental conditions Cas9WT is more efficient at inducing an indel at the targeted locus compared to Cas9DN. This finding does not concur with an earlier report, which showed a similar efficiency for both strategies (Cas9WT and Cas9DN) in HEK293 cells [96]. This

can perhaps be attributed to the difficulty in ensuring equal co-delivery of two vectors (each expressing one guide). A dose-dependent effect has been reported to reduce the activity of the endonuclease in various studies [58, 74, 145].

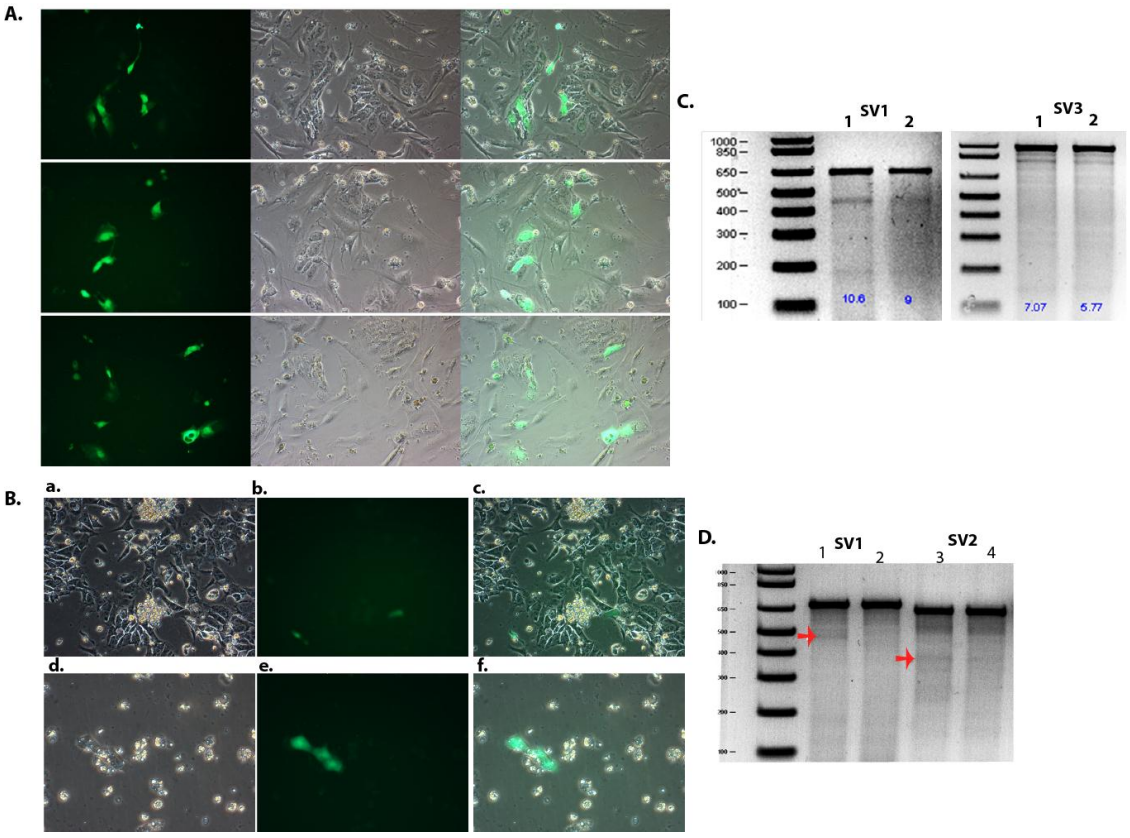

Figure 4-5: Analysis of on-target modifications following nucleofection of iPSCs with sgRNA1-Cas9-EGFP or paired Cas9n vectors. (A) Cells transfected with 2 µg of vector/2x10^6 cells. (B) Effect of plating density of transfected cells on cell survival (ECM). (C) Efficiency of on-target modification using the Surveyor assay for the cells transfected with 2 µg/2x10^6 (Lane 1) and 4 µg/2x10^6 (Lane 2) sgRNA1-Cas9WT-EGFP using two different PCR amplicons. (D) Efficiency of indels for cells transfected with sgRNA1/3-Cas9DN using two different primers (SV1 and SV2) that span the expected indel site (Lanes 1 and 3) compared to untransfected control (Lanes 2 and 4). The red arrows indicates the band location for the digested product.

In conclusion, these data indicate that different approaches can be used to transfect iPSCs with the sgRNA-Cas9-expressing plasmid. Although lipid-based transfection of plasmid DNA has been used in iPSCs [180], in this study this approach was not highly successful in these difficult-to-transfect cells [94]. Nucleofection has been reported to have a superior efficiency of plasmid delivery in pluripotent cells [94]. Although a comprehensive optimisation of both systems is needed for a proper comparison, the data from this study support the idea that nucleofection is slightly more efficient. More importantly, nucleofection is an approach that has been validated in maintaining

stem cell properties. Therefore, nucleofection was selected for the subsequent gene targeting experiments.

4.2.3 Gene correction of disease-causing mutation in iPSCs

Having established conditions for transfection of iPSCs with a beta thalassaemia mutation (BT-iPSCs), the efficiency of HDR using different design strategies to prevent further modification of the targeted allele brought about by the continuous activity of Cas9 was assessed. The effect of placing the cut site at a different location relative to the mutation or insertion site for the selection cassette on HDR efficiency was also examined using different guide pairs.

4.2.3.1 *Strategy for gene correction*

Although strategies for site-specific repair in the absence of drug selection are available, such as those using ssODN templates [78, 202], they can be limited by the low frequency of targeting of certain genomic loci and low transfection efficiency or cloning inefficiency in some iPSCs lines. Enrichment by FACS [78] or by the combination of droplet digital PCR and sib selection methods [203] has been suggested to facilitate the identification of transfected cells or those with on-target DNA modifications. However, such an enrichment strategy is complicated by the need to cryopreserve the cells in a 96-well format which is associated with a very low recovery rate in this laboratory. It was thought that efficient antibiotic selection followed by seamless excision [182] would offer a more practical alternative that avoids multiple rounds of sib selection aimed at enriching for low-frequency events [203]; this was therefore chosen as the method of choice.

The rate of HDR can be improved by using a donor plasmid with isogenic sequence and increasing the length of the homology arms [204, 205]. Therefore, isogenic sequences to those of parental beta thalassaemia-iPSCs were used to synthesise homology arms (Genscript). Three base substitutions that include two SNPs in addition to the correction of the disease-causing mutation, were introduced into the donor template sequence (Figure 4-6). The first synonymous SNP was introduced in the PAM sequence by converting the NGG sequence into NCG for one guide of the first pair (sgRNA1/3) [59]. A change in the PAM sequence was surmised to eliminate the recognition site for one guide (sgRNA1) and prevent modification in the repair template as closely spaced DNA strand breaks cannot be efficiently generated (Figure 4-7). Although the second guide of the pair is still active, the generation of a single nick has been reported to be highly inefficient in inducing NHEJ in pluripotent stem cells [58]. This was attributed to the activation of the BER mechanism [97, 111], although data from a recent study challenged this notion [145]. To assess if this exonic modification had a functional consequence, the variant effect predictor (VEP, Ensemble) analysis

tool was used to exclude any possible consequences of this variant introduction on gene function or mRNA processesing.

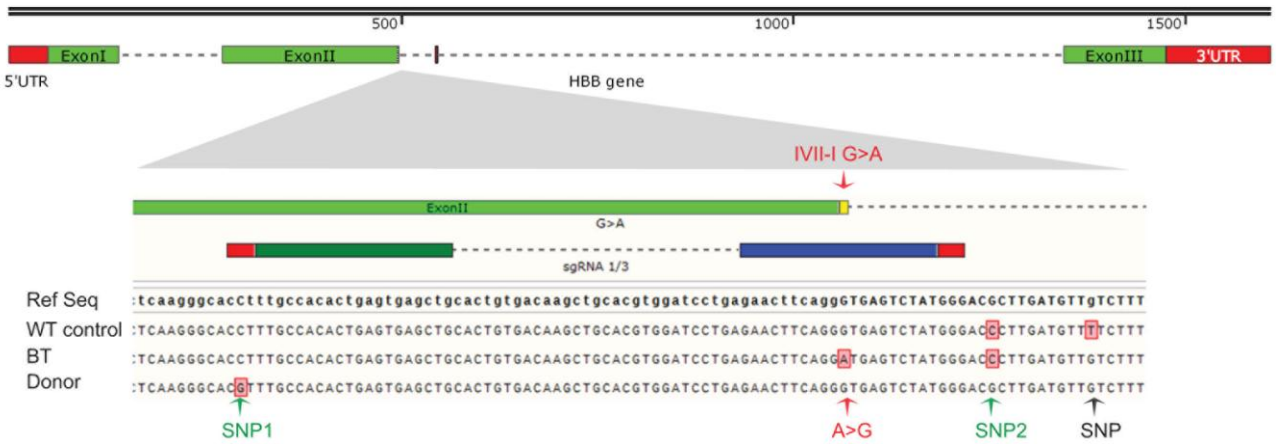

Figure 4-6: Schematic of the location of the SNPs introduced in the piggyBac donor template. Three SNPs were introduced in the donor template; (1) one in the PAM sequence of sgRNA1 to prevent Cas9 from nicking the donor template or the corrected allele when sgRNA1/3 is used (SNP1, green); (2) a polymorphic SNP variant present in the RefSeq to validate that the modification occurred through gene correction (SNP2, green); (3) and the substitution of the disease-causing mutation (A>G, red arrow). Black arrow indicate a polymorphic SNP variant.

A similar strategy was thought to be irrelevant for the second pair as a selection cassette was inserted at one of the guide-binding sites, which theoretically should reduce the efficiency DSB generation (Figure 4-7). Inserting the selection cassette at the sgRNA4 binding site would result in four consecutive 5'mismatches to the guide sequence at the targeted site. A second SNP in the intron region downstream of the mutation site (a common polymorphic SNP variant absent in beta thalassaemia-iPSCs) was also introduced.

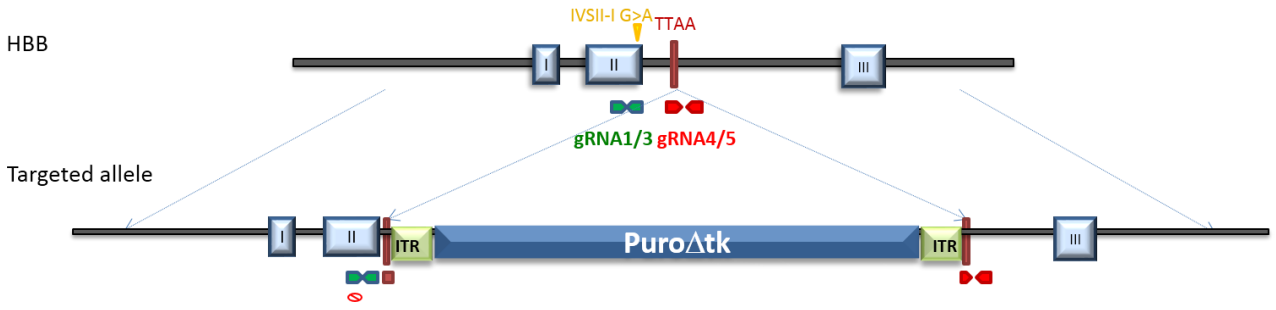

Figure 4-7: Schematic of the strategies used to prevent the formation of closely spaced SSBs. The PAM sequence was inactivated in the first pair (green pentagon) in the donor template by a SNP, while the selection cassette will destroy the first guide in the second pair (red pentagon).

The efficiency of HDR using a double nicking strategy can also be affected by other factors, such as the distance of the nicks relative to the insertion site of the selection cassette or the gene correction site (alteration site) [96]. Therefore, two different targeting strategies to correct the disease-causing mutation were assessed. Two pairs of sgRNA with the highest scores reflecting lowest likelihood of off-target activity (CRISPR design tool [58]) were selected after functional validation in HEK293FT. One guide (sgRNA3) in the first pair (sgRNA1/3) was designed to direct Cas9n to induce a nick 6 bp downstream of the mutation site whereas in the second pair (sgRNA4/5) one guide (sgRNA4) was designed to induce a single DNA break 9 bp upstream from the insertion site (TTAA) of the selection cassette and 38 bp downstream from the mutation site. The distance between the mutation and insertion sites is 47 bp (Figure 4-7).

4.2.3.2 *Optimisation of gene targeting in iPSCs*

Four independent experiments were performed using sgRNA1/3 and sgRNA4/5, as illustrated in Table 4-1. Two different forms of donor vector were used: circular and linear. In the initial experiment (A), a handful of GFP-positive cells were observed under fluorescence microscopy indicating a low transfection efficiency. The low number of GFP-positive cells observed may have been due to the decreased amount of the GFP-expressing vectors used in the transfections as the total mass of the plasmid was equally divided between the two sgRNA-expressing vectors (GFP- and puromycin-expressing vectors). Increasing the total amount of the sgRNA-expressing vectors to 3 μg did not, however, improve the fraction of observable GFP-positive cells. A Surveyor assay was carried out to verify if the transfection was associated with nickase activity for one of the pairs (sgRNA1/3, Table 4-1.A1). Faint bands of the expected fragment size were observed indicating the presence of a possible indel at the guide binding sites, which is consistent with our initial optimisation experiments (Figure 4-5.D). Again, it was not practical to calculate the cutting efficiency due to the high background and low sensitivity of the assay. Given these findings, it was thought that increasing the number of transfected cells would increase the GFP-positive proportion of cells. Therefore, $8x10^6$ cells were transfected with a total of 8 μg pDNA expressing sgRNA4/5-Cas9DN (Table 4-1.C1). However, even at a high cell density, GFP-positive cells were still rare. Unexpectedly, the number of resistant clones obtained after two weeks of positive selection for targeted events using puromycin did not correlate with the number of observed GFP-positive cells. In fact, there were 39 and 15 clones each for cells initially transfected with sgRNA4/5 and sgRNA1/3, respectively (Table 4-1.A6 and A5).

The overall low targeting efficiency in iPSCs is in agreement with Merkle *et al.* ($0.75x10^{-5}$ and $1.9.5x10^{-5}$ for sgRNA1/3 and sgRNA4/5, respectively) [145]. Too many resistant clones to count

were obtained from the high cell density transfection (Table 4-1.C1), despite the disproportionate number of observed GFP-positive cells. No resistant clones were observed for cells transfected with either the donor vector alone or a plasmid expressing non-specific sgRNA-Cas9n with a donor vector indicating that both a donor and a DNA cleaving components are required to induce a gene targeting events. Importantly, 7.5-fold more resistant clones were generated by using the circular form of the targeting vector compared to the linear form (2 clones only) (Table 4-1).

The presence of a targeting event in the absence of a proportionate number of GFP-positive cells may indicate that the fluorescence signal intensity (low GFP) was undiscernable from background (autofluorescence) by microscopy, the CBh promoter driving the expression of Cas9-EGFP is too weak in iPSCs, and/or may indicate a proportion of targeted events were induced by a single nickase activity (Cas9n-puromycin).

.

Table 4-1: Experimental design and number of puromycin-resistant clones observed

	Paired guide used	**Donor vector**		**Resistant clones**	**Comments**
A	$2x10^6$ cells/2 µg vector, 500 ng donor				Few GFP+ cells
1	sgRNA 1/3	-	-	-	Control/indel efficiency calculation
2	Control sgRNA-Cas9n	+	Circular	0	HR/non-specific integration
3	-	+	Circular	0	HR
4	sgRNA 1/3	+	Linear	2	Compare form of targeting vector
5	sgRNA 1/3	+	Circular	15	
6	sgRNA 4/5	+	Circular	39	Compare different pairs of guide
B	$1x10^6$ cells/500 ng donor				Few GFP+ cells
1	sgRNA 1/3	+	Circular	4	pDNA: 2 µg
2	sgRNA 1/3	+	Circular	6	pDNA: 3 µg
3	sgRNA 4/5	+	Circular	16	
C	3µg each sgRNA-Cas9n + 2 µg donor				Few GFP+ cells
1	sgRNA 4/5	+	Circular	Too many to count	

4.2.3.3 *Genotyping of puromycin-resistant clones*

4.2.3.3.1 Agarose gel-based genotyping

To investigate the efficiency of various strategies for mediating HR events, a total of 60 puromycin-resistant clones were selected from three different transfection experiments (A4-6, B1 and C, Table 4-1) to screen for correctly targeted alleles. Two clones from patient iPSCs transfected with sgRNA1/3 and a linear donor vector, 17 clones with sgRNA1/3 and a circular donor, and 41 clones with sgRNA4/5 and a circular donor were included in the analysis. Multiplex PCR was used to screen for the presence of the targeted allele and/or the original alleles. The screening was carried out using two sets of primers for both 5′ and 3′ homology arms as suggested in Yusa *et al.* [182]. Each primer set included a forward primer annealing outside the targeted homology arms and two

reverse primers, one located across the insertion site and the other at either the 5′ and 3′ ITR of the transposon (insertion specific primer) (Figure 4-8.A).

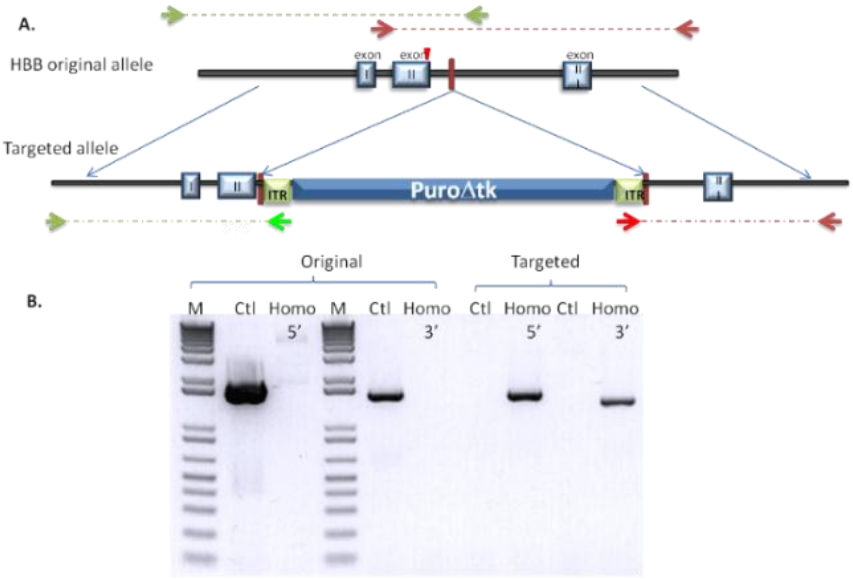

Figure 4-8: Screening strategy for the correctly targeted allele and validation. (A) Primer sets used to screen the 5′ (green arrows) and the 3′ regions (red arrows) of the homology arms. (B) Validation of each set of primers using BT-iPSCs and a clone targeted at both alleles (after confirmation by sequencing).

This multiplex PCR strategy facilitated rapid genotyping of the clones by enabling the simultaneous screening for the presence of the original and/or the targeted allele. Each primer pair set was validated for the amplification of the proper fragment size using either the BT-iPSCs or homozygous clones identified during the screen as positive controls and verified by sequencing (Figure 4-8.B).

Ten clones were excluded due to poor growth characteristics or difficulties in maintenance. Cell lysates were used for the screening of the remaining 50 clones [182]. No band for the targeted allele was observed in clones generated with a linear donor indicating that their drug resistance was acquired through random integration of the selection cassette (Figure 4-9). Almost half (52.08 per cent) of the clones obtained using a circular donor vector had one targeted allele (heterozygotes), and 16.67 per cent had a possible targeting event at both alleles (possible homozygous or hemizygous) after excluding clones with random integration (31.26 per cent) (Details provided in the next section; Figure 4-10 and Table 4-2). The probability of obtaining a targeting event at one or two alleles was dependent on the pair of guides used ($p=0.011$, Fisher's exact test). Thus, this suggests that the efficiency of HDR can be influenced by the efficiency of the guide pair, the location of the nick sites relative to the insertion site or by sequence-dependent factors. Of note, the

analysis (Fisher's exact test) was performed by including all the targeted events (with or without random integration).

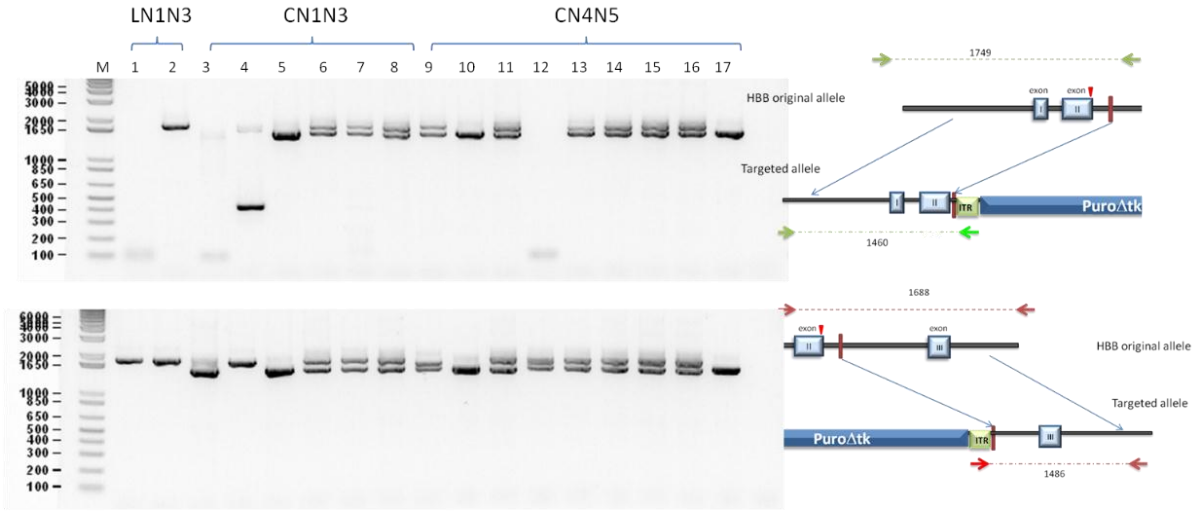

Figure 4-9: Screening of puromycin-resistant clones by multiplex PCR at both left (top gel) and right (bottom gel) arms. The band sizes for amplicons generated by untargeted and targeted alleles differ by 200–300 bp, which facilitates genotyping. Lanes 5, 10 and 17: clones targeted at both alleles, Lanes 6, 7, 8, 9, 11, 13, 14, 15, 16: clones targeted at one allele; Lanes 3, 12: targeting at one arm, while Lane 4 represents large deletions at one arm. L: linear donor; C: circular donor; N1, 3, 4, 5: sgRNA(N_AN_B)-Cas9n.

Of the 48 clones targeted with circular donor vector, four clones showed a targeting event in the right arm, coupled with no amplicons in the left arm (Figure 4-9: Lane3, and Lane 12) which may indicates a deletion occurring at this site that abolished the primer binding site. Furthermore, an irregular band pattern, in terms of band size, was observed for amplicons obtained from the original allele, which may indicate an on-target indel-induced by the Cas9DN at this allele. For instance, the band pattern for the right arm (bottom panel, Figure 4-9) in Lane 9 indicates that the clone may contain a large deletion in the original allele. Overall, 81.5 per cent of the resistant clones (regardless of random integration) included in the screening had at least one targeted allele (52.08+16.67+6.25+6.25), which indicates that this strategy can enables efficient gene targeting in iPSCs.

4.2.3.3.2 Screening for the presence of random integrations

The screening for random integration was carried out using a PCR-based strategy. The primer pairs were designed as suggested by Yusa *et al.* [182] where each primer set is designed to bind to a DNA region that spans the junction site between the homology arms (left and right arms) and the plasmid backbone (Figure 4-10.A). Both a negative (beta thalassaemia-iPSCs) and a positive control (donor plasmid) were used to validate the assay. A total of 33/48 (25 heterozygous, and

eight homozygous) clones were identified that were devoid of random integrations. Interestingly, the screen revealed a higher number of random integrations associated with sgRNA1/3 (42.8 per cent) compared to sgRNA4/5 (20.59 per cent) (Table 4-2). However, this difference was not statistically significant (p=0.145, Fisher's exact test) likely indicating that the process is not directly associated with the sgRNA-Cas9DN activity.

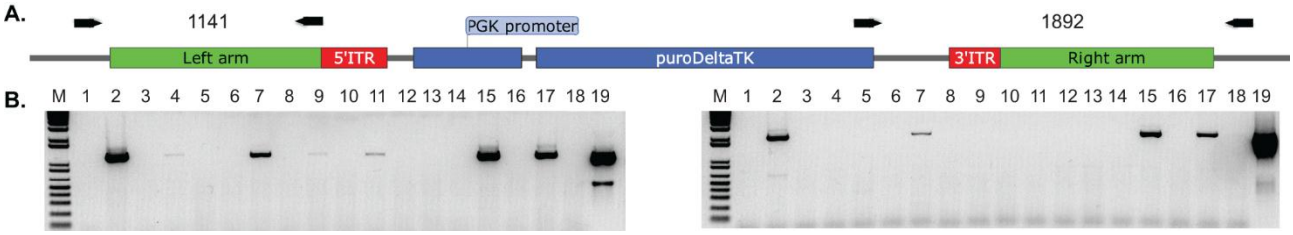

Figure 4-10: PCR-based screening for random integration for both left and right arms using whole cell lysate. (A) Schematic indicating the binding sites of primer pairs used in screening for random integration at both left and right arms. (B) Analysis of amplicons obtained by PCR using 1.5 per cent agarose gel. Bands in Lanes 2, 7, 15 and 17 indicate the random integration of both arms, Lanes 4, 9, 11 indicate integration for one arm. Lane 19: donor plasmid used as positive control.

Interestingly, some clones carried a random integration for one arm of the vector (e.g., Lane 11), while others showed integration at both arms (e.g., Lane 2) (Figure 4-10.B). The observed difference in the band intensity can be attributed to the difference in the DNA amount as crude cell lysate was used for the analysis, or a copy number of integrated transgenes. All positive clones (random integration) were excluded from further analysis.

Table 4-2: Frequency of clones with correctly targeted alleles in puromycin-resistant clones

		Screened								**Excluded**	**Total**
		Excluded					Included targeted allele		Total screened		
Guide pair	Donor form	R	M	R+M	HT+R	HO+R	HT	HO		Growth characteristics	
N1N3	Linear		2						2		2
N1N3	Circular	3	1	2	1	-	6	1	14	3	17
%		21.43	7.14	14.29	7.14	-	42.86	7.14			
N4N5	Circular	2	1	-	2	3	19	7	34	7	41
%		5.88	2.94	-	5.88	8.82	55.88	20.59			
Total	Circular	5	2	2	3	3	25	8	48		
%		10.42	4.17	4.17	6.25	6.25	52.08	16.67			
Total	L+C	17					33		50	10	60

R: Random integration; M: different band size (compared to control), indicating either deletion or insertion or targeting at one arm, HT: one allele targeted, HO: both alleles targeted, L: linear, and C: Circular.

4.2.3.3.3 Sequence verification of selected clones

After excluding clones with random integration, the sequence of both the left and right arms of some selected clones was verified by Sanger sequencing (Figure 4-11.A and B). The sequence analysis of the left arm confirmed the correction of the disease-causing mutation and the introduction of the intended two SNPs upstream and downstream of the correction site indicating that gene modification occurred through HR (Figure 4-11.A). Notably, no on-target modification was detected at the DNA binding site for either sgRNA1 or sgRNA3. Surprisingly however, sequencing of the right arm revealed the presence of a deletion at the binding site for one of the guides in one clone. Evidently, separating the guide by insertion of the selection cassette did not completely eliminate indel formation at the targeted site. A recent study demonstrated that the on-target indel frequency was reduced to 0.5 per cent ±0.4 per cent upon physical separation of the binding site of the guide pair with an insert in conjunction with Cas9DN [145]. The latter value was obtained by screening 188 clones at eight targeted sites. It was reasoned that on-target indels would be found to be a rare occurrence if a wider screen with a larger number of targeted clones was carried out (see below). Nevertheless, the presence of DNA modification at one guide binding site (sgRNA5) suggests that DNA nicks can be repaired through mutagenic pathways.

Notably, the identified clone with the on-target mutagenesis was a compound heterozygote with different deletions occurring at both the targeted allele and the non-targeted allele (Figure 4-11.C). Sequencing of the original allele in heterozygotic clones (n=3) confirmed the earlier observation of abnormal band patterns at the original allele and revealed the presence of a DNA modifications in all the sequenced clones (n=3) (Figure 4-12). Merkle *et al.* [145] reported a high frequency of on-target modifications at the untargeted allele, with a mean value of indel frequency of 26 per cent using the Cas9DN strategy. Therefore, it was postulated that it would be possible to identify clones with no indel at the original allele if more clones were screened. This is supported by the observation that sequencing revealed an intact original allele in two monoallelically targeted clones that were excluded because of the presence of random integration of the donor vector. Consequently, all the clones identified with monoallelic targeting (n=25) were screened for the presence of indel at the original (or untargeted allele) in addition to both left and right arm DNA regions of the guide binding sites at the targeted allele.

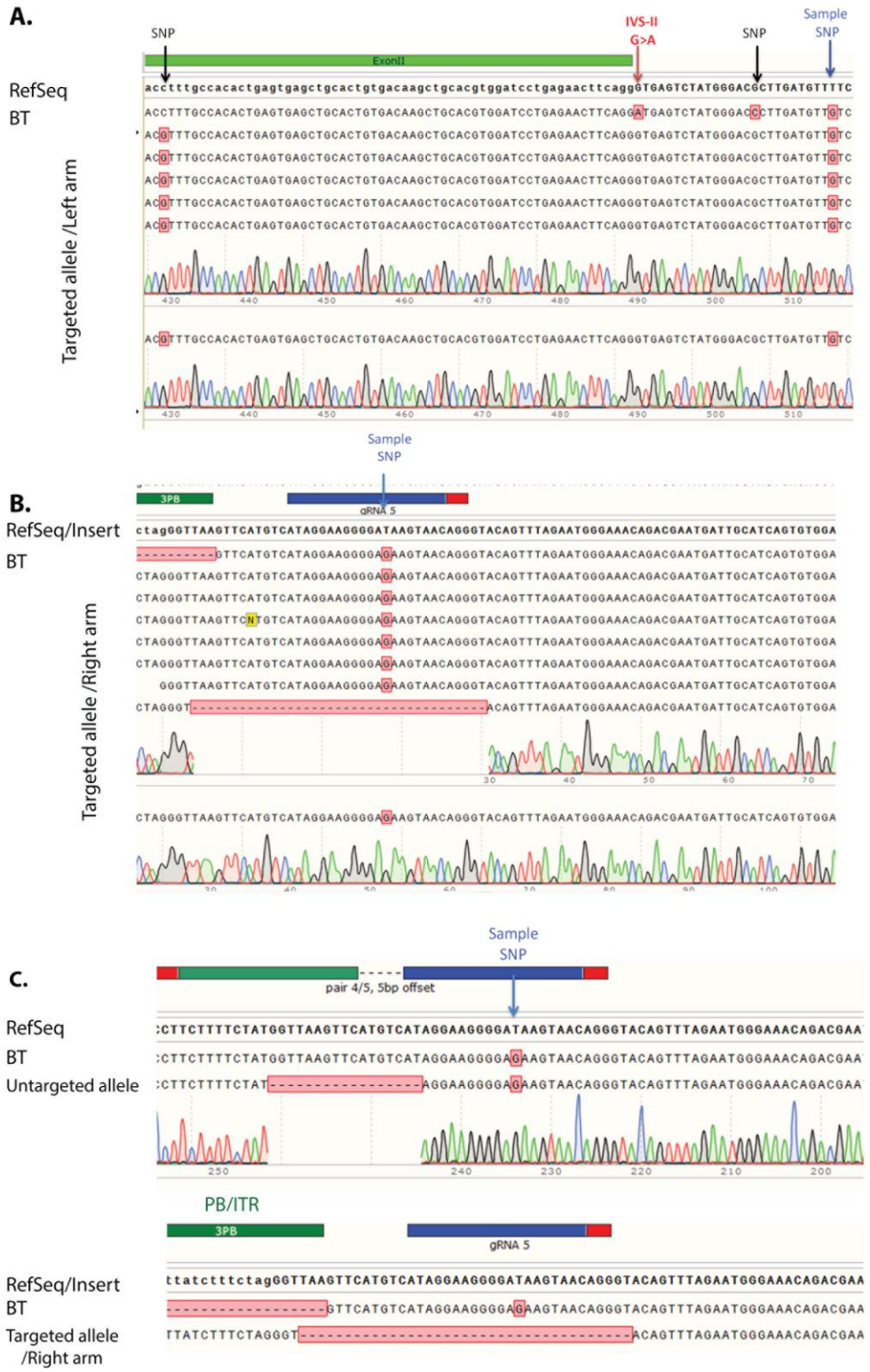

Figure 4-11: Sequencing of the DNA region around the targeted site in the corrected clones identified by PCR screening. (A) Sequencing of the left arm of the targeted allele revealed the correction of the disease-causing mutation and the presence of both SNPs. (B) Sequencing of the right arm identified the presence of a deletion at the gRNA5 binding site in one clone. (C) Sequencing of the original allele for the clone with possible nickase activity in (B) revealed the presence of an indel at both the targeted and the untargeted alleles. PB/ITR: *piggyBac* inverted repeat region. BT: beta thalassemia-iPSCs.

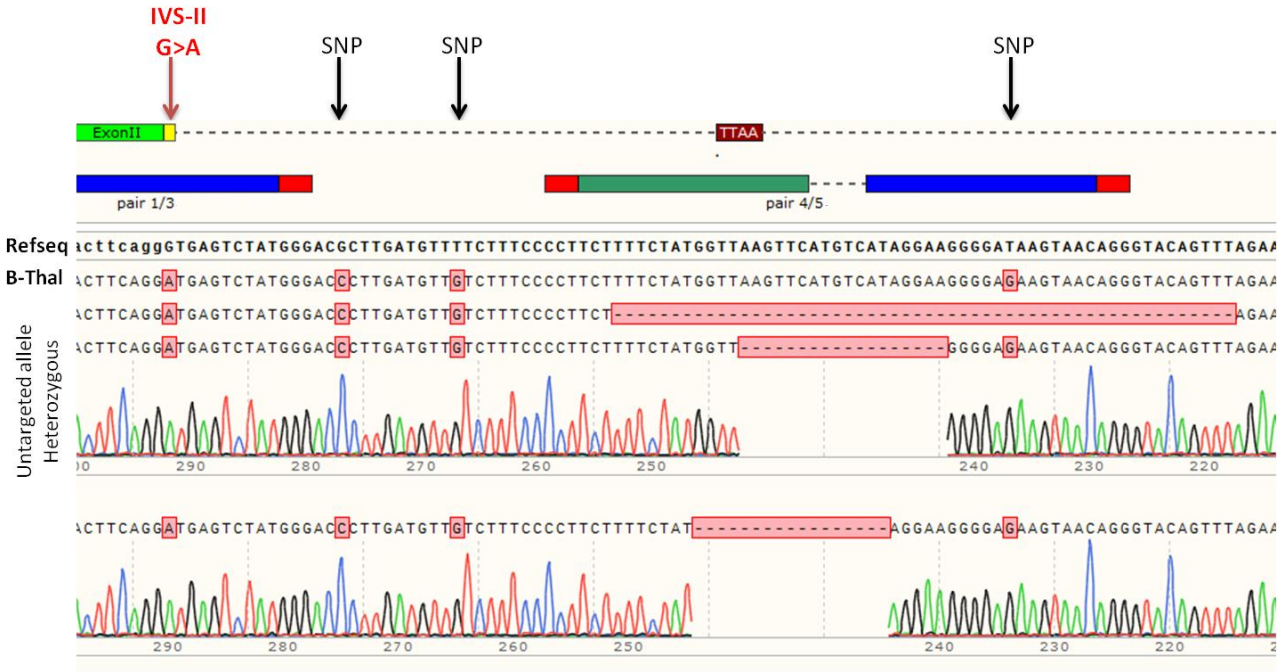

Figure 4-12: Sequencing of the original allele in clones targeted at one allele revealed the presence of indels in each of the selected clones.

Screening the untargeted allele will provide a more comprehensive analysis of the on target mutagenesis at the original allele as a result of the activity of either dual or single Cas9n. Screening the targeted allele will specifically assess the frequency of indel generation as a result of nickase activity, as one of the guide pairs is either inactivated (due a mutation in the PAM motif) or destroyed by the insertion of the selection cassette. To assess this, the HMA was recently put forward as an efficient high-throughput method for screening for Cas9-induced allelic alterations [183]. A genotype screen was therefore carried out using HMA and the results verified by sequencing random samples.

4.2.3.3.4 PAGE-based genotyping (HMA) of nickase binding sites

PAGE-based genotyping was used to further characterise on-target alterations mediated by either nickase or paired nickase. Three sets of primer pairs that spanned the predicted indel sites were used to amplify short amplicons around the Cas9n binding site at both arms of the targeted insertion and in the original allele. Each primer set included a forward gene specific primer coupled with either a reverse gene-specific (untargeted allele) or insertion specific (targeted allele) primer. To screen for alteration at either the original or the targeted allele, PCR amplicons generated from the BT-iPSCs and the donor plasmid (or one homozygous targeted clone verified by sequencing) were used for the formation of DNA duplexes, respectively. Heteroduplex mobility is slower due to the formation of the open angle at the site of the mismatch [183], resulting in a different band migration patterns in

native PAGE for homoduplex and heteroduplexes as well as for heteroduplexes with a different type of mutation mismatch. The variability in the band pattern was assessed relative to a control homoduplex formed by hybridisation of PCR product from the control sample (e.g. BT-iPSCs) to rule out artefacts produced during PCR amplification.

(A) On-target modification at the original allele in heterozygous clones (due to Cas9n or Cas9DN activity):

The large variability in the size of PCR amplicons across clones was discernible by agarose gel (Figure 4-13.A). However, HMA revealed the complex composition of the PCR products obtained from a single clone. In fact, multiple heteroduplex bands were observed for some clones (Figure 4-13.B and C: Lane2, 18 and 19), indicating that those clones were heterogeneous. The heterogeneity may have resulted from a possible mosaicism due to non-unicellular source, or may indicate modification in a subset of cells as a result of continuous nuclease activity as suggested in a previous study [145]. The data also indicate that indels were present at the untargeted allele in all 25 potential heterozygous clones included in the screen. Notably, the high rate of indels at the untargeted allele was irrespective of the guide pairs (sgRNA1/3 and sgRNA4/5) used to target the locus. Moreover, PAGE-based analysis also showed that different types of mutations had been introduced as revealed by the distinct, and unique band patterns observed for each clone (Figure 4-13.B, C and D).

Three samples were randomly selected for sequencing. Sequence analysis provided further evidence that indels occurred at the sgRNA binding sites with complex allelic alteration patterns that included insertions, deletions and substitutions on the same allele (Figure 4-13.E, and F). Remarkably, the introduction of a single base substitution (G>A) was observed at position 16 of the binding site for one of the guide (sgRNA5) in two out of the three sequenced clones (Figure 4-13.E and F: L4 and L17). To achieve a more comprehensive understanding of the mutation pattern induced by Cas9-DN strategy, additional heterozygous clones were sequenced. Sequencing data revealed a mutation spectrum that consisted of a deletion range that varied between 5 and 54 bp, affecting the intervening sequence (offset sequence) between the guide binding sites (Figure 4-13.G and H, and Figure 4-14.A). The latter was observed in all the untargeted alleles analysed (Figure 4-14.A).

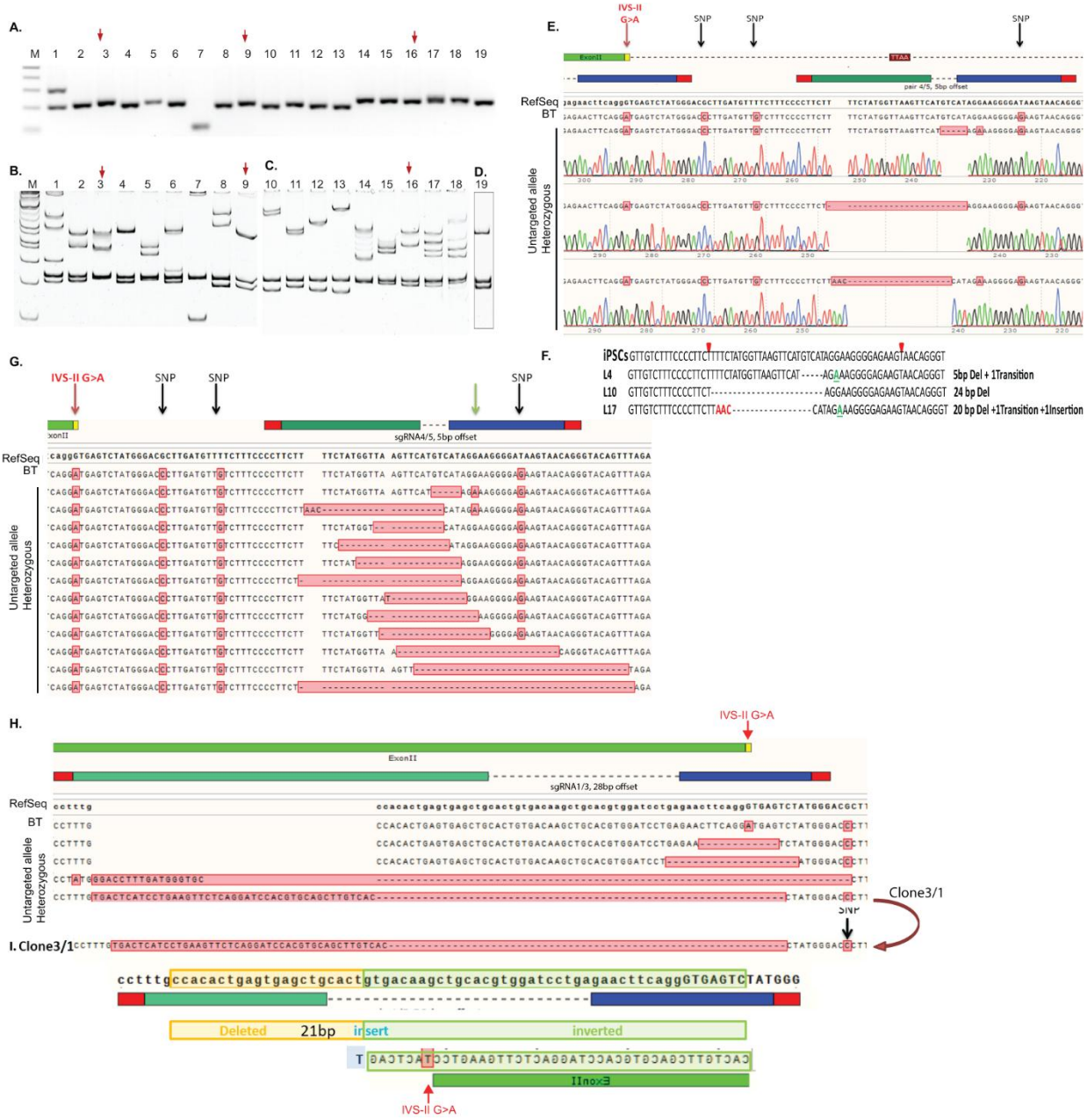

Figure 4-13: Detection of undesired mutagenesis at the original allele in heterozygous clones. (A) Band patterns of PCR products obtained by amplifying a short DNA region that spans the guide RNA binding site, resolved on 2 per cent agarose gel. (B, C, D) PAGE-based genotyping of the original allele in heterozygotes clones revealed a Cas9-DN- or Cas9n-mediated indel at the original allele in all heterozygote clones screened. (E and F) Validation of the HMA results by sequencing three randomly selected clones (Lanes 4, 10 and 17) obtained by sgRNA4/5 and the mutation pattern observed. (F) The red triangle indicates the nick site, dashed line: deletions, red bases: insertions and green bases: substitutions. (G) Sequencing result of additional clones (9) targeted by sgRNA4/5. (H) Sequence result of untargeted alleles in four clones obtained by targeting with sgRNA1/3. The sequence revealed a pattern that includes small deletions, deletion + insertion and deletion + inversion + insertion. (I) analysis of the complex mutation (deletion, insertion, and insertion) observed in one clone (C3/1).

The active guide binding site (minimum 17 bp [92]) remained intact in 7/12 and 2/4 alleles of cells targeted with sgRNA4/5 or sgRNA1/3, respectively. Short microhomology regions (2-4) flanking the deleted sequence were identified in 67 per cent of the clones verified by sequencing (sgRNA4/5) (Figure 4-14.A and B) [206]. Usually, this pattern of deletion is associated with the C-NHEJ DNA repair pathway [110].

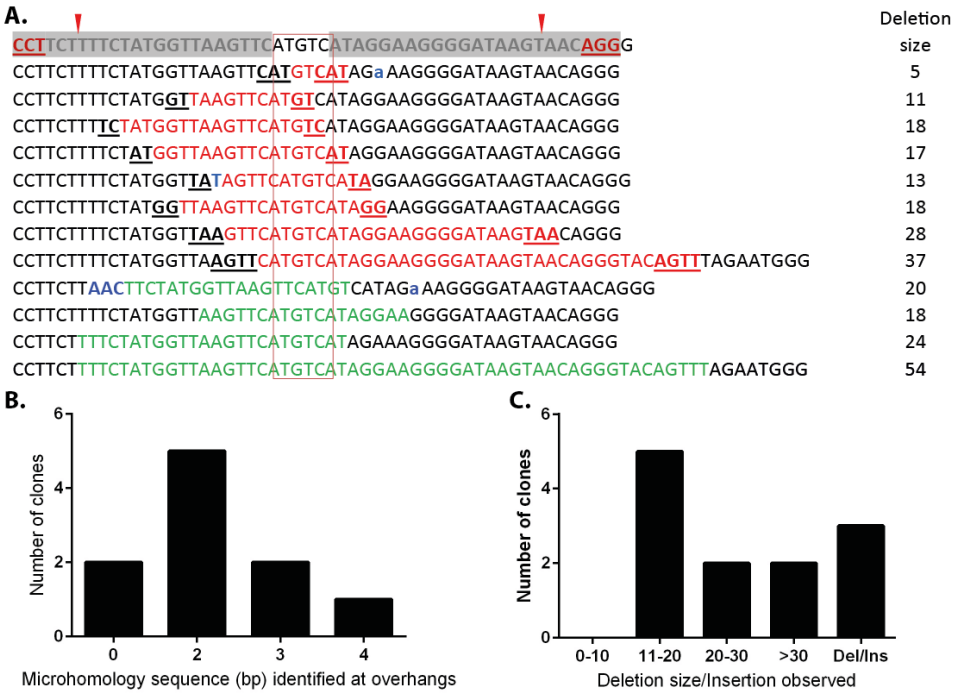

Figure 4-14: Mutation patterns observed at the original allele in heterozygous clones targeted with sgRNA4/5-Cas9DN. Micro-homology sequence flanking the deleted region varied between 2–4 bp in 8/12 clones. (B) the observed frequency of clones with the indicated micro-homology sequence size. (C) The mutation pattern included deletions or mixed (deletions, insertions, and substitutions), the frequency of clones identified with the indicated size range is shown. sgRNA sequences are highlighted in grey flanked by the PAM sequence in red, micro-homology sequences are underlined, deleted sequence flanked by micro-homology are in red, and the deleted sequences in the remaining clones (4) are in green, insertion is in blue, and substitutions in lower case.

Intriguingly, an inversion with deletion and an insertion of 17 bp was detected in two clones targeted with sgRNA1/3 (Figure 4-13.H and I). These data suggest that the length of the overhang may play a role in determining the mechanism of repair. In fact, different mutation patterns have been reported for ZFNs and TALENs, which were also attributed to the differences in the spacer size [207]. However, the small number of clones analysed for sgRNA1/3 guide pairs does not allow proper and reliable analysis. The repair speed and the local DNA environment can also affect the repair path [131]. The targeted site for sgRNA1 may be located in a region of nucleosome occupancy (PSC track not available) which may impact the speed of repair, resulting in a different

mutation pattern compared to that of sgRNA4/5 (Figure 4-15). Interestingly, nucleosome positioning has recently been implicated in restricting Cas9 activity [80].

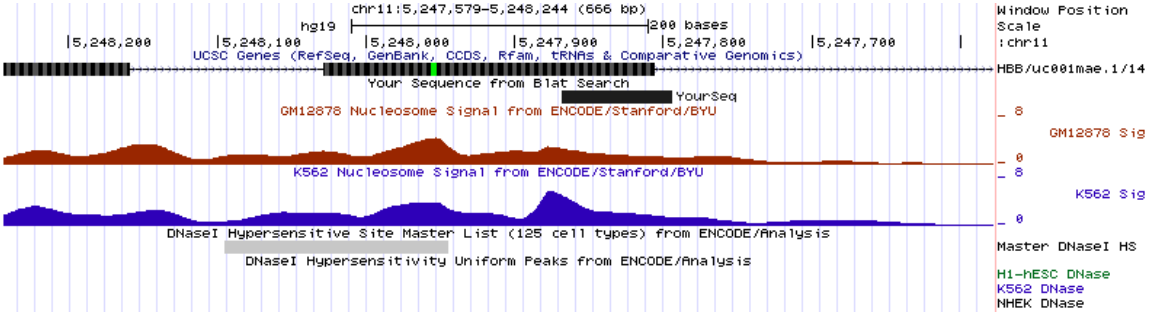

Figure 4-15: Nucleosome occupancy for the DNA region targeted by sgRNA1/3-Cas9DN in two different cell lines. The targeted region for sgRNA1/3 is shown as a black solid box.

The unexpectedly high rate of on-target mutagenesis observed at the untargeted allele can perhaps be attributed to the amount of plasmid used as current evidence suggests that increasing the plasmid dose can increase the on-target mutation frequency [74, 145]. Indeed, an on-target indel frequency of 26 per cent at the original allele was reported by Merkle *et al.* [145] upon using 1 μg of the vector (3-6 x10^6 cells) whereas 2 μg (2x10^6 cells) was used in the current study, resulting in 25/25 clones with monoallelic targeted events that carried a mutation. However, it should be noted that despite the difference in the amount of plasmid used, the majority of the targeted clones (20/25) did not carry multiple alleles, as judged from their band patterns and confirmed by sequencing (for limited clones), whereas Merkle *et al.*[145] identified multiple alleles possibly as a result of continuous plasmid expression which was reasoned to lead to the formation of multiple types of indels in a subset of cells within the same clone. These data can be interpreted as meaning that while on-target mutation can certainly be influenced by the vector dosage, locus-dependent characteristics can also play a role in modulating this activity.

(B) On-target modification at the guide binding site in both left and right arms (Cas9n activity)

No mutation were detected at either the sgRNA3 or the sgRNA4 binding sites. Remarkably, the on-target allelic modification was identified at sgRNA5 in the right arms of heterozygote (Ht) and homozygous (Ho) clones (3/19 Ht and 2/6 Ho) (Figure 4-16.A: Lane 4, 12, 18 and C: Lane 3 and 7) targeted with sgRNA4/5, despite the separation of the guide pair. This indicates that indels most likely resulted from the activity of a single nickase.

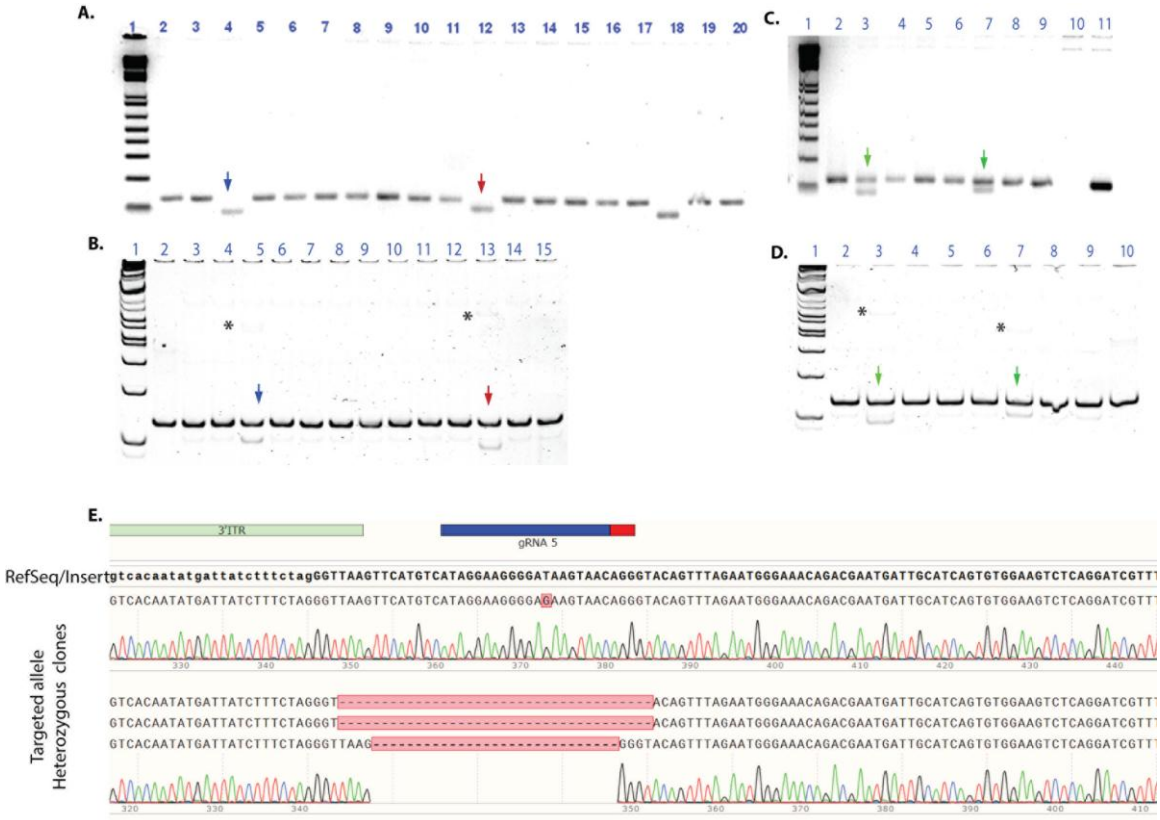

Figure 4-16: Screening for DNA modification at the guide binding site on the right arm in DNA obtained from targeted clones. PCR amplicons of (A) 19 clones with one targeted allele (Ht) (C) or in clones with biallelic targeting (Ho) were separated on 2 per cent agarose gels. Three bands with deletions were observed in Ht and two in Ho clones. (B, C) HMA confirmed the presence of the deletion. Arrows with similar colours indicate the same clones and asterisks indicate the heteroduplex band location. Hybridisation was carried using unpurified PCR products. (E) Sequence analysis of the sgRNA5 binding site for identified clones (Lanes 4, 12 and 18 in Panel A). Lane11: homozygous clone verified by sequencing (biallelic targeted clone: BT-T).

Table 4-3: Frequency of on-target mutagenesis observed at the sgRNA5 binding site in mono- and bi-allelic targeted clones obtained using a strategy that destroy the binding site for one guide of the pair sgRNA4/5 .

sgRN4/5						
Original allele			No. of clones analysed	Indel	Clone intact at original allele	
	Heterozygous		19	19	0	
Targeted allele			**No. of clones analysed**	**Indel**	**No. of targeted alleles**	**Indel allelic frequency at the targeted allele**
	Heterozygous	L/arm	19	0	19	15.79%
		R/Arm		3		
	Homozygous	L/Arm	7	0	14	14.29%
		R/Arm		2		
	mean					15.04%

In contrast to the findings of Merkle *et al.*[145], these observations suggest that this design strategy can also result in a relatively high frequency of on-target mutagenesis as a result of nickase activity associated with one of the guide (gRNA5). In this study, the on-target indels induced by sgRNA5-Cas9n at the targeted allele with a mean allelic indel frequency of 15.04 percent (Table 4-3).

The presence of indels was further confirmed using HMA (Figure 4-16.B and D). The faint band observed for the heteroduplex bands may be the result of the disproportionate amount of PCR product used to form the mix. In fact the intensity of the band for the control (homozygous clone verified by sequencing) (Figure 4-16.C: Lane 11) indicates a higher PCR efficiency for the control clone compared to other tested samples (the intensity was more than two-fold higher, as calculated by Image Lab software). This highlights the importance of using normalised amounts in the mix if one wishes to reliably detect the heteroduplex.

On the other hand, the analysis of bands for the left arms revealed an abnormal band pattern in only one clone which suggests a possible deletion at one allele and an insertion at the second allele (Figure 4-17.B: a lane2). However, a similar band pattern was also observed at the original allele (Figure 4-13.A: Lane 2), which may indicate a duplication or a heterogeneous clone (clone excluded from further analysis). Although it was possible to identify six correctly biallelic targeted clones from eight homozygous clones (for both sgRNA1/3 and sgRNA4/5), all the heterozygous clones carried an indel at the untargeted allele.

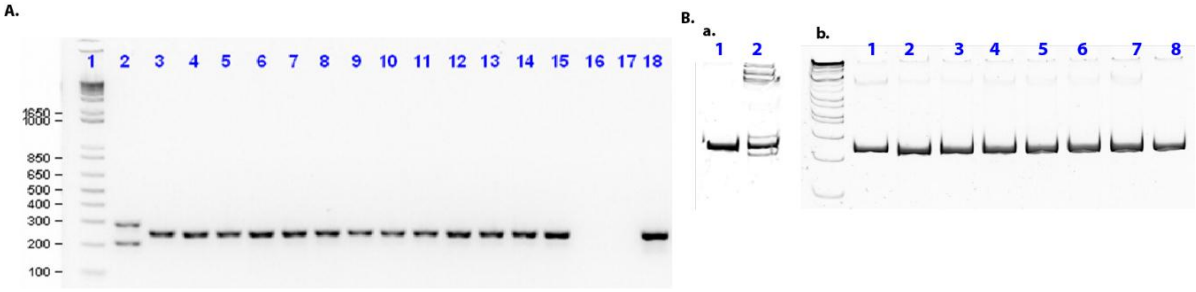

Figure 4-17: Screening for DNA modifications at the guide RNA binding site on the left arm of DNA extracted from targeted clones. (A) PCR band patterns resolved on 2 per cent agaros gels indicating that the clone in Lane 2 has both a deletion and insertion at this site, Lane 2 2–15: targeted clones, Lane 18: positive control (donor plasmid), Lane 16: negative control (BT-iPSCs) and Lane 17: PCR blank control. (B) PAGE-based genotyping indicating the absence of heteroduplexes with the exception of the sample in B (a) Lane 2 (same clone in A, Lane 2).

Overall, these data suggest that nickase activity could also be associated with mutagenesis, possibly by single break conversion to DSBs with subsequent stimulation of the repair mechanism through the error-prone NHEJ pathway [208]. The majority of the reported studies on nickase activity were performed in immortalised or transformed cell lines such as HEK293, K562, Hela and U2OS cells [58, 61, 96, 98, 209], and described an activity that varied between minimal to absent, whereas in

PSCs the activity was reported to be undetectable [58]. However, a more recent study observed an indel frequency of 4 per cent at the guide binding site at the original allele (untargeted allele in targeted clones) induced by a monomeric nickase activity in iPSCs [145]. The mean calculated allelic frequency in the current study is 15.04 per cent associated with one guide only (sgRNA5), while no mutagenesis was associated with sgRNA3 or gRNA4. Although the indel frequency in iPSCs seems to be higher than that reported by either Tsai *et al.* [98] of 1.02 per cent (0.0048 per cent to 3.04 per cent) or Hsu *et al.* [58] of 1.1 per cent in transformed cell lines, the indel frequency in both the current study and that of Merkle *et al.* [145] was not estimated in the bulk cell population. In fact, in the present study, the double nickase strategy that displayed a 100 per cent indel frequency in the targeted clones was below the detection limit of the Surveyor assay in the bulk cell population.

Upon further examination of the data reported by Merkle *et al.* [145], the frequency of indels induced by Cas9n varied across different loci and over a wide range under the same experimental conditions (plasmid-based expression, amount of plasmid used, ratio to donor template, transfection conditions etc), at the targeted allele (mean=4 per cent). This may indicate that the sequence context of the targeted site can be a determinant of either the repair mechanism [131] induced by the DNA nick or the activity of the guide. The latter has been demonstrated to be affected by the sequence content [210]. Furthermore, although the Merkle *et al.* [145] study reported a total mean indel frequency for Cas9WT that is higher than Cas9n (10-fold difference for the untargeted allele), the distribution pattern of the data was similar. In other words, both the Cas9WT and the nickase variant can have high and low frequencies of mutagenesis depending on the locus, which was particularly obvious at the targeted allele. Remarkably, the average frequency of mutagenesis at one locus (HCRT-C2) was higher for Cas9n (compared to Cas9WT).

It is likely that the relatively high rate of on-target indels (at the targeted and untargeted allele) in the current study can be attributed to a combination of factors in the strategy used. For example, these factors include the use of plasmids for transfection, as plasmids have been demonstrate to produce sustained expression of Cas9 protein over several days, which can expose cells to continuous activity of the enzyme, as opposed to using Cas9 protein, which is cleared from cells in less than 24 hours [93, 94]. Moreover, the amount of plasmid used for transfection can affect the efficiency of endonuclease activity [58, 74], while chromatin factors at the targeted locus, the sequence of the targeted site and the donor template strategy [145] have been reported to affect the targeting efficiency and activity of the Cas9 protein. A recent report described a loss of blue fluorescent protein (BFP) signal, reflecting a NHEJ event, from a BFP reporter upon transfection of cells (HEK293) with the ribonucleoprotein form of Cas9N and ssDNA donor [71]. Similar

observations under different experimental conditions may indicate that mutagenesis with nickase is a consistent finding that can be obtained with Cas9N delivered either as a plasmid [145] or as a ribonucleoprotein [71] at certain loci. Whatever the precise mechanism, nickase is clearly associated with mutagenesis at a rate that is sequence- and/or locus-dependent [145]. The dual nickase approach was suggested to maintain the targeting efficiency of Cas9WT while reducing the potential for adverse mutagenic off-target activity [61, 96]. This was based on the hypothesis that SSBs are mostly resolved through the efficient housekeeping process that removes SSBs [97]. However, the results of the present study shows that undesired on-target indels can be avoided for some guide whereas the presence of an indel at the binding site of another guide binding site suggests that nickase can use an error-prone repair mechanism, which casts doubt on the suitability of this approach for producing safe, precisely edited iPSCs destined for clinical applications. A better understanding of the factors and conditions that determine the choice of repair path following DNA nicks is needed to design foolproof strategies that permit efficient and specific gene targeting in iPSCs.

4.2.4 Assessment of off-target activity for gRNA

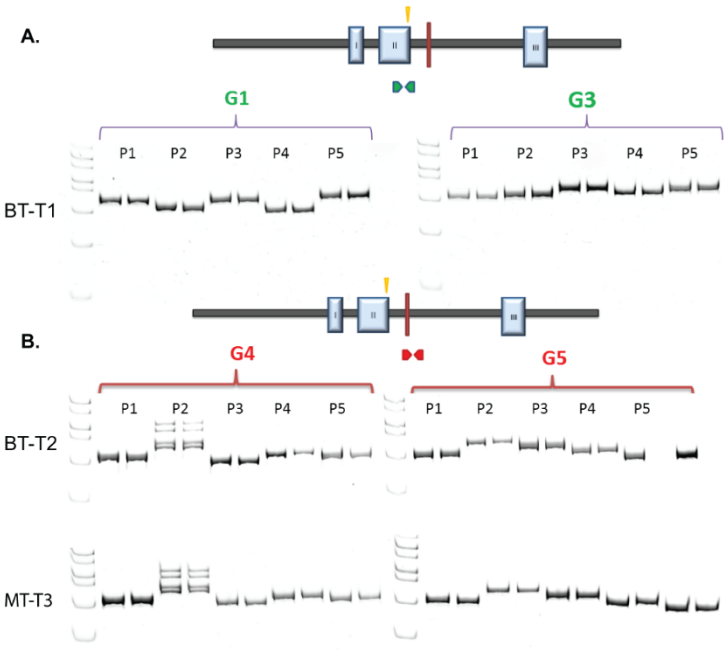

Figure 4-18: Screening of five potential off-target loci for each gRNA in three different targeted clones. Cropped image of PAGE gels for hybridised PCR products obtained by mixing amplicons from the parental clone (BT-iPSCs) and the indicated targeted clone using five pairs of primers. Primers were used to amplify the DNA fragment across the potential off-target site (P1–5) for clones transfected with either (A) sgRNA1/3 or (B) sgRNA4/5. Two bands are shown for each site; the first is for the hybridised products of BT-iPSCs (control), and the second is for the hybridised products obtained after an equimolar mix of PCR amplicons derived from BT-iPSCs with the targeted clone.

To examine the possibility of off-target mutagenesis induced by Cas9n, the five top-scored genomic sites that bore high sequence homology to the sgRNA-targeted region were screened for the presence of mutations. Ten potential DNA sites for each guide pair (five sites per sgRNA) were screened using 10 different primer pairs that span the intended genomic site in three different clones, one generated using the sgRNA1/3 and two using the sgRNA4/5 guide pairs. PCR amplicons obtained from the selected clones were mixed with amplicons obtained from the BT-iPSCs, and the hybridised product was analysed using native PAGE [183]. The presence of heteroduplex bands was used to indicate the presence of allelic mismatches (relative to the BT-iPSCs parental line) at the screened DNA site. No off-target activity at the selected loci was detected as evident from the presence of only a single band, indicating that only homoduplexes were formed between amplicons (Figure 4-18).

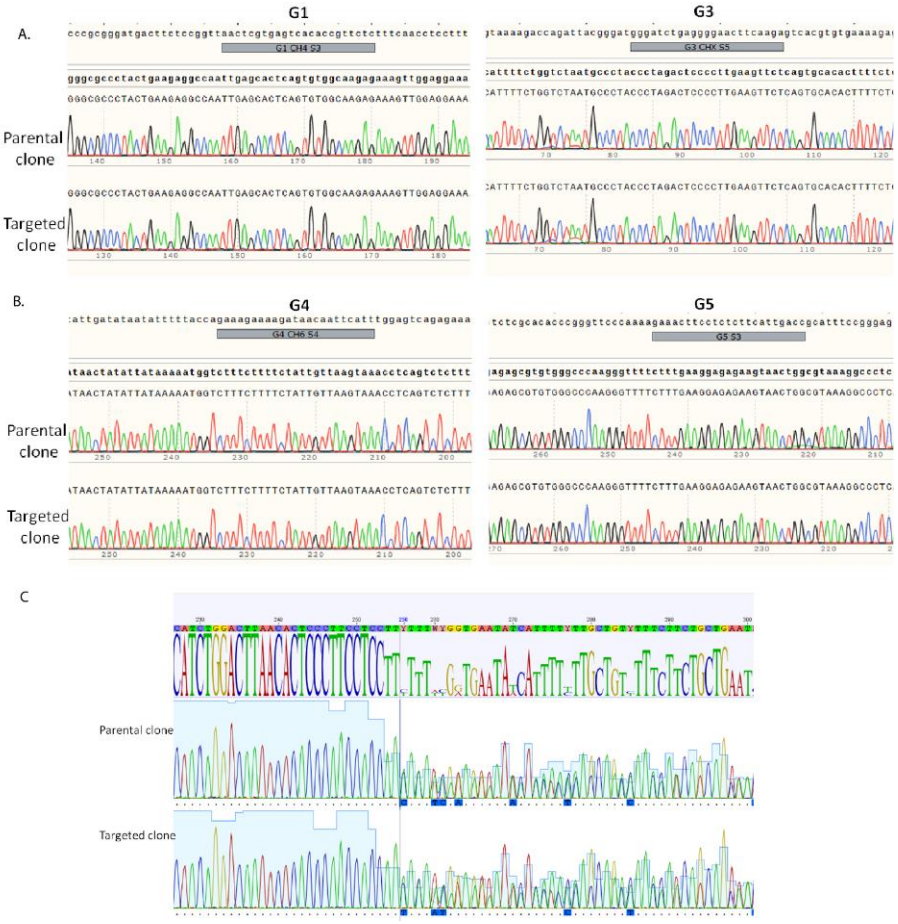

Figure 4-19: Verification of selected off-target sites by Sanger sequencing. A representative sequence of one off-target site for (A) sgRNA1/3 and (B) sgRNA4/5, showing the sequence in the parental clone (BT-iPSCs, upper) and that of the targeted clones (lower). (C) The sequence at the P2 site for potential off-target sgRNA4 indicates possible heterozygosity present in the both the parental (BT-iPSCs) and targeted clones.

The presence of four heteroduplexes and the absence of a homoduplex at one site (P2) for sgRNA4 in both screened clones does not suggest off-target activity, as a similar pattern was observed in BT-iPS clones. Rather, the pattern suggests heterozygosity at this locus.

The sequence was further verified by Sanger sequencing for three randomly selected off-target sites (P1-5) for each gRNA, including the DNA site for off-target 2/sgRNA4 (Figure 4-19.A and B). Sequence analysis confirmed the heterozygosity of the BT-iPS clones at this locus and the absence of DNA modifications in the targeted clones (Figure 4-19.C). The sequence was further deconvoluted using Surveyor mutation software by SoftGenetics (Pennsylvania, USA). This finding highlights the importance of including the parental line as a control to detect heterozygosity at the analysed DNA sites. The presence of unique SNPs in some cell lines raises potential concerns about the limited reliability of the currently available *in silico* predictive tools that identify potential off-target sites based on homology to the consensus genomic sequence.

4.2.5 Seamless excision of the selection cassette in the corrected clones

4.2.5.1 *Excision of the selection cassette*

The main rationale for using the *piggyBac* transposon system in gene correction is that it affords eventual seamless excision of the selection cassette upon transfection with transposase. Enrichment of cells that have lost the expression of the viral thymidine kinase (truncated version, ΔHSV-TK in this case) marker can be achieved through negative selection with 1-(2-deoxy-2-fluoro-beta-D-arabinofuranosyl)-5-iodouracil (FIAU) [182]. In the presence of FIAU, a nucleoside analogue, the active expression of Δtk driven by PGK leads to production of a toxic metabolite that kills the cells that express the gene. In order to establish an optimum dose of FIAU suitable for negative selection of the corrected iPSCs with a restored allele, a dose response curve was generated using biallelically-targeted cells. Interestingly, a high background was observed over a wide range of FIAU concentration (0.2-50 μM) using positive control cells (BT-T: homozygous targeted cells) plated either on either MEF or feeder free on ECM. The low efficiency of the FIAU selection possibly indicates that the cassette was silenced, perhaps because it is inserted into an inactive locus (*HBB*) not expressed in iPSCs.

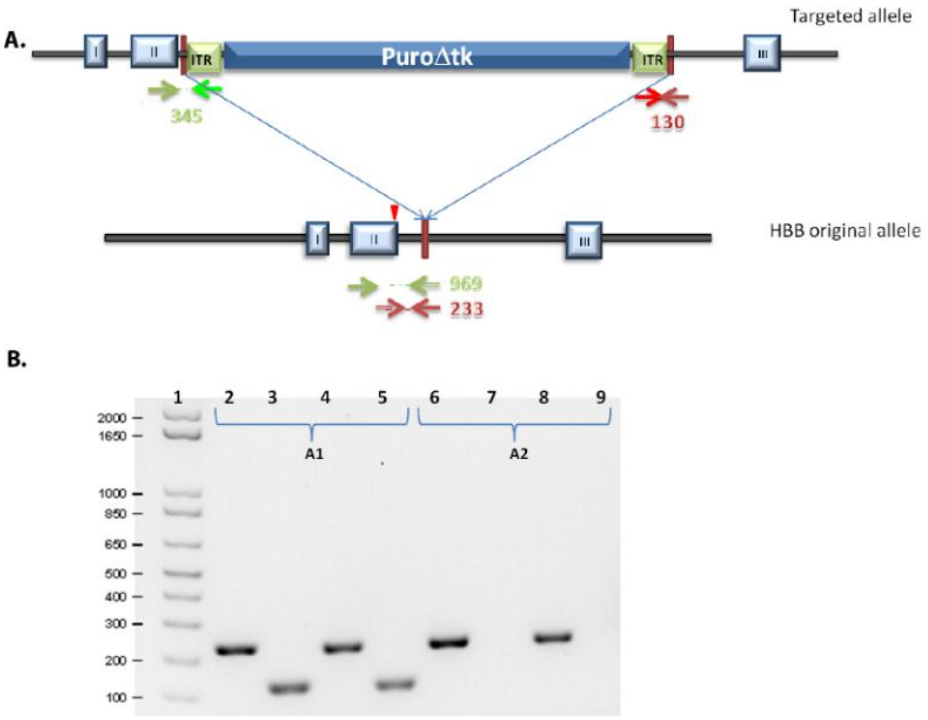

Figure 4-20: Schematic of the screening strategy used to identify gene-corrected clones with excision of the selection cassette insert and its validation. (A) Two sets of multiplex primers were assessed. Each primer set include three different primers: (1) a common gene-specific primer upstream or downstream of the insertion site (→←), (2) insertion-specific primers (→←) and (3) a reverse gene-specific primer (→←). This strategy allows simultaneous amplification of the corrected allele that excised the selection cassette and the targeted allele if present. (B) Amplicons obtained by using the gene-specific primer pair (red arrows) in the presence (A1) and absence (A2) of the insertion-specific primer P3 (ITR region of *piggyBac*) in iPSC lines with a beta thalassaemia mutation (BT-iPSCs: Lanes 2 and 6), WT-iPSCs (Lanes 4 and 8) and two biallelic targeted line (BT-T iPSCs: Lanes 3, 5, 7 and 9).

A similar observation was reported in a study where a different (hygromycin) selection cassette driven by PGK promoter was inserted into the first intron of the *HBB* gene in iPSCs. These authors attributed the decrease in transgene expression to epigenetic factors after expansion of the cells in the absence of selection for a few passages [141]. Despite these caveats, we set out to test if the low efficiency of the FIAU selection would be sufficient for the enrichment of the restored allele in the corrected clones. Two biallelically-targeted (A, and B) and two monoallelically-targeted (C, and D) clones, were initially transfected with a plasmid expressing HA-tagged hyperactive *piggyBac* transposase. The expression of the transposase was driven by a CMV promoter. Transfected cells were passaged on Day 3 (in the absence of puromycin) and plated at low cell density before subjecting them to FIAU selection for eight days. Clone screening was carried out using a set of primers that included a common primer located upstream or downstream of the excision site and two additional primers, either a gene-specific or an insertion-specific primer (Table 2-5.A and B).

Each primer set was designed to generate amplicons with different sizes to indicate the presence or absence of the selection cassette (Figure 4-20.A) [182].

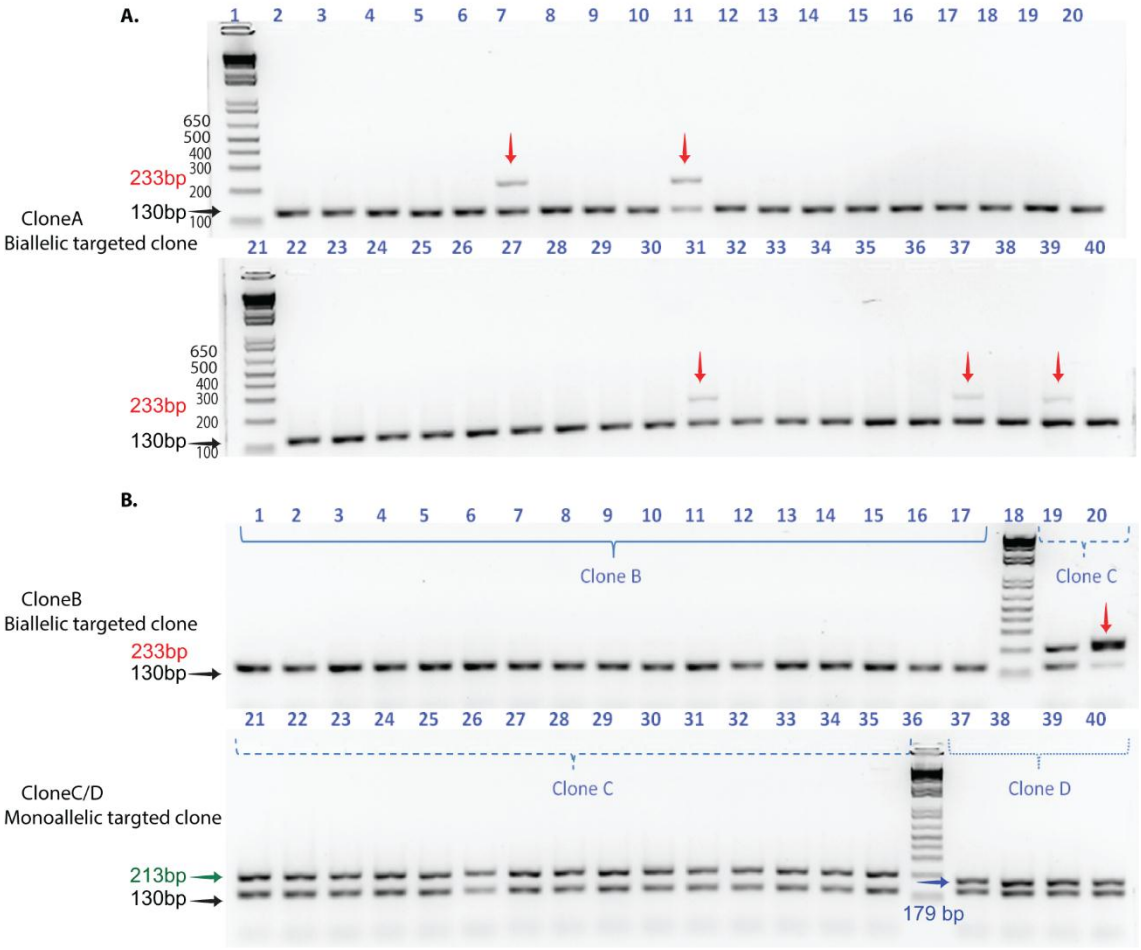

Figure 4-21: Screening for the restored allele following transposase transfection and selection with FIAU. Band patterns obtained using the screening strategy shown in Figure 4-20 upon transfecting (A) a biallelic targeted clone with transposase, followed by FIAU selection. The pattern indicate that the excision of the selection cassette occurred in a small subset of cells of the identified clone (Lanes 7, 11, 31, 37 and 39); (B) whereas none was detected in either clone B (biallelic targeted clone) or clone D (monoallelic targeted clone) and one in clone C (monoallelic targeted clone, B: Lane 20). PCR band size of 233 bp (red arrows) indicate a restored allele and PCR band size of 130 bp (black arrows) detect the presence of the insert. The size of the amplicons generated by the untargeted allele in the monoallelic targeted clones (MT-T) displayed deletions that were different from that obtained for the restored allele (green and blue arrows). The deletions resulted in either a small shift in the band (clone C, 213 bp) or a different band size (clone D, 179 bp).

The band pattern was used to genotype the DNA region spanning the TTAA site. The screening strategy was validated using the BT-iPSCs and cells carrying a biallelic integration of the cassette (Figure 4-20.B). Eight clones derived from clone A and one from clone C (from 48 screened clones) displayed a band pattern that indicated possible excision of the cassette, whereas none was detected in clones derived from B and D (Figure 4-21). Of note, the size of the amplicons generated

by the untargeted allele in the heterozygous clones displayed deletions that were different from that obtained for the restored allele. The deletions resulted in either a small shift in the band (clone C) or a different band size (clone D). To overcome this hurdle an alternative strategy for enriching and isolating iPSCs with rare events was adopted. The approach was adapted from McCormick *et al.*[211] and Miyaoka *et al.* [203] where the former proposed enrichment by fractionation with a simultaneous use of an assay to detect the presence of the modified gene, while the latter applied the approach in conjunction of droplet digital PCR in iPSCs [203, 211]. Therefore, in order to avoid screening of hundreds of cells, the applicability of using a stepwise approach that involved: (1) FIAU selection to eliminate cells expressing Δtk (inefficient), (2) a sequential fractionation of the positive cell pools that displayed a positive band (screening with PCR) for the restored allele to enrich for cell populations with positive events, (3) plating cells at low density with subsequent screening and isolation of clones with the restored allele by limiting dilution (details are provided in the Material and Method) was assessed. The PCR strategy used to detect excised events (positive cells) is shown in (Figure 4-20.A).

Cells from two different biallelic targeted clones (E, and F) were transfected and passaged on Day3 for two cycles (instead of one), with subsequent plating at low cell density into two different plates (pool 1 and 2) before subjecting them to FIAU selection for eight days [182]. The cells were dissociated and divided into two fractions. The first fraction was plated at low density and the second fraction was used for DNA extraction and screened using multiplex PCR (Figure 4-22). Two bands were observed in both fractions indicating the presence of alleles with an excised cassette.

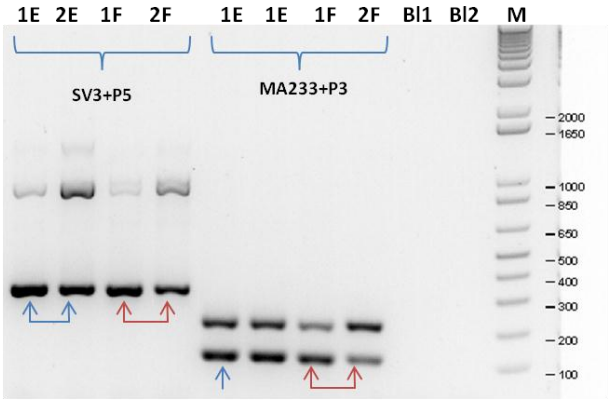

Figure 4-22: Screening of pools of iPSCs to identify the fraction with the highest percentage of alleles with an excised selection cassette. Screening for cassette excision was carried out using two primer sets (SV3+P5 and MA233FR+P3) in cell pools collected at Passage 2 after transfection with transposase and subjected to FIAU selection for eight days. Two cell pools from each clone were screened (Clone E: 1E and 2E, Clone F: 1F and 2F). The pool with the highest proportion of excised events was selected for further screening after plating cells at low density. Bl1 and Bl2 are blank controls for the different primer sets.

The fraction with the highest percentage of the restored allele (Figure 4-22: Lane 2E and 2F) was used for subsequent culture. The variable band intensity between the different primer sets most likely reflects differences in primer efficiencies. Nevertheless, both primer sets indicated that the restoration of the original corrected allele occurred in a subset of cells. Clones obtained from the low-density plating were further screened for biallelic excision. A mixed band pattern was observed in the majority of the clones (Figure 4-23.A and B).

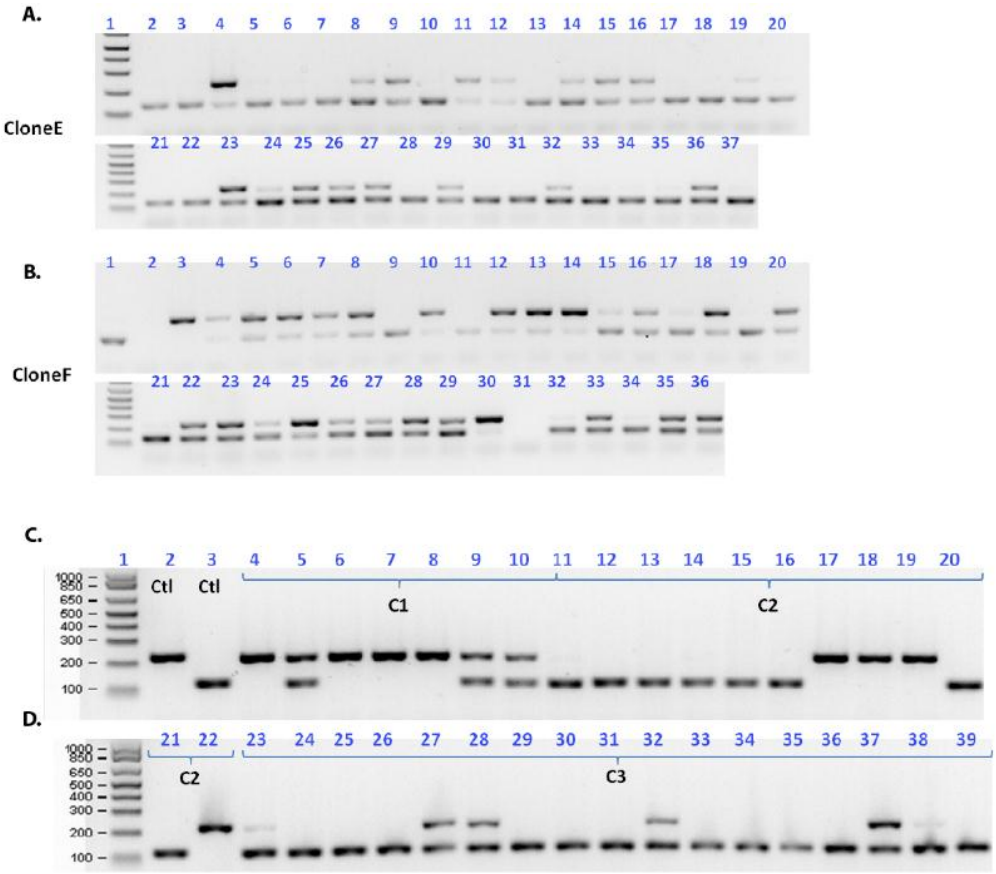

Figure 4-23: Screening of clones for biallelic excision of the selection cassette. (A and B) Band patterns observed upon screening clones obtained by plating cells with the highest proportion of excised events for two clones with biallelic targeting. (C and D) Band patterns observed upon screening clones obtained by deriving single cell-based clones by limiting dilution in a 96-well plate. The C1 clones were derived from Clone E, Lane 23, C2 clones were derived from Clone F, Lane 25 and C3 were derived from Clone E, Lane 8.

Most of the clones were either negative for the excision or contained heterogeneous cells, as indicated by the disproportionate band intensity of the amplicons. Satisfyingly, few clones displayed a biallelic excision in the bulk cells population (Figure 4-23.A: clone E: Lane 4 and clone F: Lane 3 and 30) as indicated by the prominent band observed for the restored allele. Consequently, clonal isolation by limiting dilution was performed for selected clones followed by a

second round of screening (Figure 4-23.C and D). Two cycles of plating enriched cells followed by screening enabled us the purification and isolation of clones with biallelic excision as indicated by the presence of a single band. Remarkably, while the band pattern for some clones indicated that they mostly formed from cells with either biallelic excision or no excision (Figure 4-23: C, D: clone C2), other clones displayed a pattern that indicates monoallelic excision Figure 4-23: C, D: clone C3). It should be noted that some wells contained more than one clone which explains the disproportionate intensity of bands in some lanes (Figure 4-23: Lane 23, 28 and 32).

4.2.5.2 *Screening for random reintegration of the cassette*

Although transposase allows precise excision, the enzyme activity can be associated with reintegration of the excised insert into the genome. To assess this, the clones were screened for the presence of reinsertion of the cassette using a primer pair that detects the presence of the puromycin insert. Some clones (e.g., 5/7 for colonies derived from one parental clone: BT-T iPSCs) displayed a puromycin-specific band, while others did not (Figure 4-24.A). The sequence of the amplicon obtained from the gene-corrected clones confirmed the precise biallelic restoration of the TTAA site after excision (Figure 4-25) and an absence of cassette reinsertion was confirmed by puromycin resistance gene-targeting PCR (Figure 4-24.C).

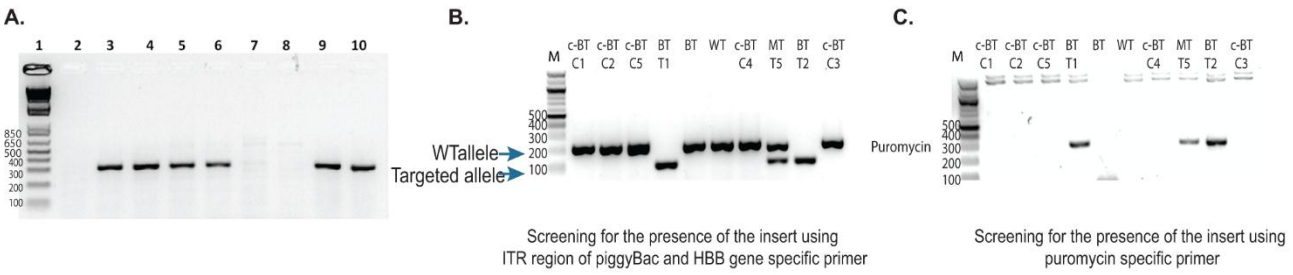

Figure 4-24: Screening for the excision of the selection cassette and the absence of reinsertion. (A) Amplicons obtained from colonies derived by limiting dilution from the same iPS clone indicate the absence of reinsertion in 2/7 clones screened (Lanes 7 and 8); Lane 2, negative control (BT-iPS clone); Lane 3, positive control (clone with targeted allele). (B) Analysis of amplicons obtained by multiplex PCR from selected clones indicates selection cassette excision in clone C1, C2, C5 derived from clones T1, T2, and T5. Analysis of amplicons obtained with puromycin resistance gene-targeting PCR indicates the absence of reinsertion of the selection cassette in biallelic (c-BT-C1 and C2) and monoallelic (c-MT-C5) gene-corrected clones following transfection with transposase and single cell cloning by limiting dilution (from original targeted iPS clones BT-T1, BT-T2, and MT-T5). BT-T1, BT-T2, and MT-T5: Parental targeted iPSCs, BT : beta thalassemia iPSCs, and WT: wild type iPSCs control.

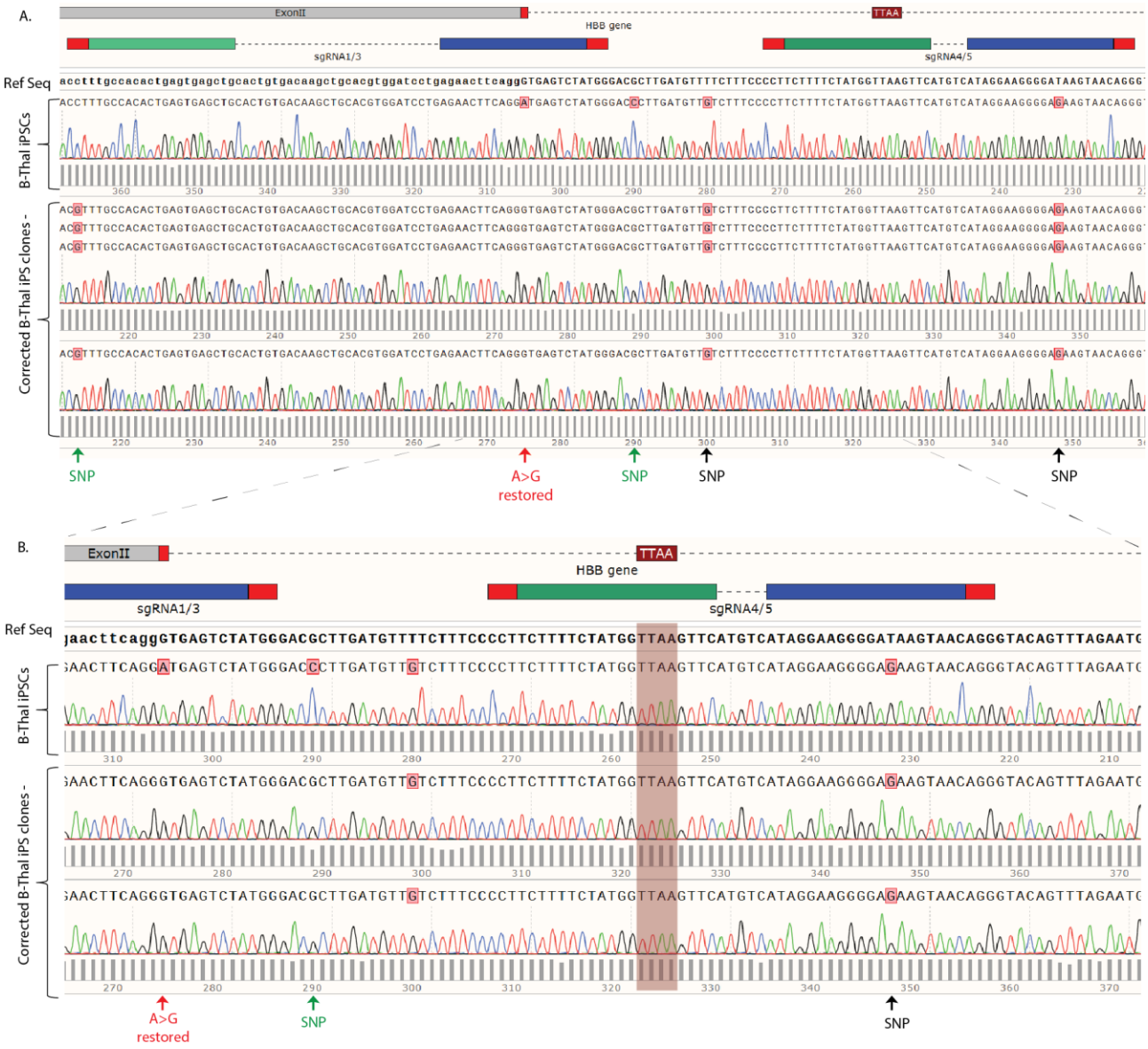

Figure 4-25: Restoration of the TTAA site after excision of the cassette. (A) The sequence of the corrected allele showing three SNPs at the designed location: a SNP in the PAM sequence, correction of the disease-causing mutation, and a SNP introduced in the intron. (B) Enlarged section of Figure (A) displaying the sequence of the mutation and TTAA sites in BT-iPSCs, and corrected iPSCs after excision of the insert..

4.2.6 Characterisation of gene-corrected iPSC clones

Three clones were selected for further experimentation and validation (gene activation). Two clones that are seamlessly corrected at both alleles and one clone that was heterozygous with one allele carrying an intron II 20bp deletion were subjected to karyotype analysis (Figure 4-26). Each of these clones displayed a normal karyotype and expressed pluripotent markers that include OCT4, and TRA-1-60 (Figure 4-26).

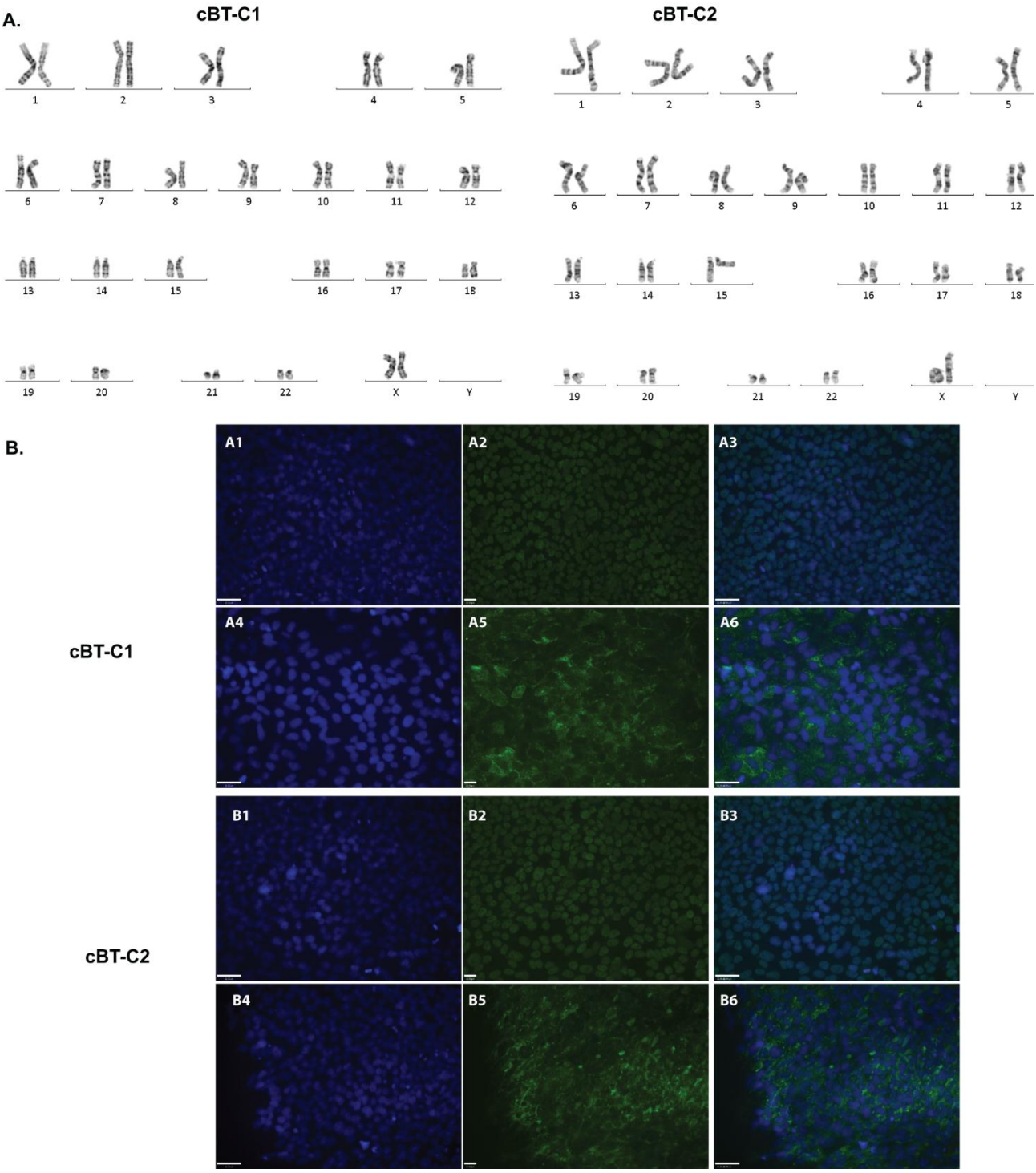

Figure 4-26: Characterization of the corrected iPSCs.

(A, and B) Representative karyotype analysis of two gene-edited iPSC lines with uniform expression of pluripotent stem cells markers that include (A1-A3, and B1-B3) OCT4 and (A4-A6, and B4-B6) TRA1-60.

4.3 *Conclusion*

This chapter describes the successful seamless site-specific gene correction of a beta thalassaemia causing mutation (IVSII-I G>A) in iPSCs using the double nickase (Cas9-DN) approach [95, 96]. In agreement with a previous study, the efficiency of the targeting events in the bulk iPSC population was low [145]; however, both biallelic and monoallelic modifications were achieved. Six out of 33 correctly biallelically targeted clones without random integration were obtained using two gRNA pairs sgRNA1/3 and sgRNA4/5. Genotyping of the guide binding site at the original allele (untargeted allele) in heterozygous clones revealed a considerable frequency of on-target mutagenesis that occurred at a rate that is higher than that reported by Merkle *et al.* [145]. It is known that the efficiency of Cas9 activity can be affected by the amount of vector used [58, 74, 145] and it may therefore be possible to isolate clones with unmodified original alleles by titrating down the sgRNA-Cas9n vector concentration. However, this will also reduce the efficiency of HR-mediated repair.

To improve the editing rate (prevent on-target activity at the targeted allele) following HR, two different design-strategies that prevent generation of the on-target DSBs in the donor vector and targeted allele were implemented [145, 198]. No on-target mutagenesis was observed for clones targeted with the guide pair with the inactivated PAM site. However, unlike the Merkle *et al.*[145] study, which suggested that separating the guide binding sites can abolish on-target mutagenesis, the result of the current study indicate that the indel allelic frequency at the targeted allele remained relatively high for one guide. It is not unlikely that the mitigating effect reported by Merkle *et al.* [145] following the separation of the nick site in the context of the DN strategy is attributable to the reduced load of sgRNA compared to Cas9n as the amount was divided between the guide pairs. This inference is based on the observation that mutagenesis was associated with monomeric Cas9n strategies (4 per cent at the original allele), whereas no indel was reported for monomeric Cas9n in the context of nick site separation in the DN strategy. Unfortunately, the Merkle and colleagues's study [145] did not provide data on the utility of this strategy with guides that display the highest on-target activity at the targeted allele.

Even though the number of guides and loci evaluated in the current study are limited, the data appear to indicate that mutagenic activity of monomeric nickase can be sequence-dependent rather than reflecting a design strategy. In other words, the paired nickase approach is not a true dimeric approach; in fact, monomeric nickase retains some activity that can be guide sequence- or locus-dependent [98, 145]. Clearly, inactivating one guide will reduce on-target mutagenesis at the targeted allele, but the rate will depend on the activity of monomeric sgRNA-Cas9n. In fact, guide sequence composition can play a role in forming stable structures that have been associated with

efficiency of activity using Cas9WT [210]. Guide 5 had the highest number of guanine bases and is likely to generate more stable DNA/Cas9n complex. Moreover, a difference in the mutation patterns was observed at the original allele for clones targeted with sgRNA1/3 and sgRNA4/5. The offset distance (28 and 5 bp) differed, generating longer overhang with sgRNA1/3 which may explain the more complex pattern of deletions, inversions and insertions. However, in contrast to Kim *et al.* [207] the larger overhang was associated with a larger insertion. Interestingly, the sgRNA1 binding site resides in a high nucleosome occupancy area, which has been reported to restrict Cas9 activity [80]. However, it is not clear if this will also impact on the mechanism of repair as the speed of repair have been reported to determine the repair path in the G2 phase of the cell cycle [131].

Despite the unexpected mutagenesis with one of the sgRNA, this strategy was effective at limiting on-target mutagenesis as evidenced by the large difference in the indel frequency observed between the targeted and untargeted alleles. Indeed, using this strategy, it was possible to identify clones with correct biallelically targeted events. Although the data did not reveal off-target activity upon screening the top predicted site based on sequence homology, nicks can be occasionally repaired through error-prone pathways, as evidenced by the presence of on target-mutagenesis [71, 98, 145].

To restore the integrity of *HBB* locus following gene targeting, the *piggyBac*-based donor design that facilitates seamless excision of the selection cassette was adopted [182]. Despite the fact that negative selection was inefficient, perhaps due to possible silencing of the cassette [141], clones with restored wild-type sequence were isolated at both alleles by implementing a stepwise approach for enrichment and screening, followed by clonal isolation by limiting dilution. This strategy was adapted from Miyaoka *et al.* [203] and McCormick [211]. HMA-based genotyping was also validated as an effective and reliable approach for screening for the on/off-target changes in genomic DNA [183]. For this assay, including a control to identify heterozygosity in the parental cell line reduced false positive results, while using an equimolar amount of the hybridised products ensured a reliable and sensitive assay in the absence of PCR product purification. Thus, with careful design of the targeting strategy and determination of the right balance between Cas9-DN activity and gRNA design, this method should be able to be used to improve editing rate for the seamless correction of any gene mutation in iPSCs, provided this is combined with careful screening and analysis of both alleles in putative corrected clones and verification of the absence of transposon-mobilised cassette reintegration.

CHAPTER V:

TARGETED TRANSCRIPTIONAL ACTIVATION OF THE β-GLOBIN GENE

5 TARGETED TRANSCRIPTIONAL ACTIVATION OF THE β-GLOBIN GENE

5.1 *Introduction*

To demonstrate the restoration of the wild type phenotype in gene-corrected disease iPSCs, the corrected cells are commonly differentiated into specialised cell types that express the edited gene [77, 212]. However, differentiated iPSCs mostly fail to adopt a mature phenotype, and this evidently limits the functional analysis of the corrected allele for genes that are associated with mature phenotypic characteristics. This is particularly relevant for correction of disease-causing mutations in the β-globin gene, such beta thalassaemia major [35, 141]. Indeed, the β-globin gene is only expressed during the late stages of erythropoiesis and is considered an adult attribute (HbA). It has been shown that erythroid cells, derived by the currently available methods for directed *in vitro* differentiation of iPSCs, mostly adopt an embryonic phenotype with variable expression of foetal globin and failure of efficient globin switching to β-globin [154]. To circumvent this problem, in this study an alternative approach was considered, which takes advantage of CRISPR-mediated transcriptional activation of the endogenous gene [156] to assess the restoration of adult gene expression following gene correction of a mutation associated with a loss of function in monogenic disease.

Using the null Cas9 variant as a targetable programmable protein to direct a single $VP16_{(n)}$ [79, 95, 163] or multiple TDs (VP64-p65-Rta and VP64 +MS2–p65–HSF1) to cis-DNA elements in the proximal promoter region of genes [165, 166] has been shown to be sufficient to modulate gene expression in a range of cell types, such as HEK293 and iPSCs. Activation of transcription through this technology is achieved by recruiting various chromatin remodelling and transcription factors to the targeted sites, which facilitates the assembly of the preinitiation complex [213, 214]. This approach can also be adapted to mediate gene activation by directing epigenetic regulators to the enhancer regions of specific loci [172]. Although variability in the potency of gene activation between different targeted gene regions was observed [160], the activation level was sufficient to enable the replacement of reprogramming factors (*OCT4*) [160] in iPSC generation or to induce the differentiation of hiPSCs [160, 166, 215]. For some loci, gene expression levels can be achieved that approach those of native tissue, provided enhanced activation systems are used [166]. Notably, the activity of some guide combinations has been reported to be cell-type dependent (under some experimental conditions). For example, guides targeting the promoter region of the *HBG1* gene increased the expression of this gene in HEK293 cells, yet failed to do so in human dermal fibroblasts [216]. In the case of epigenetic modulation of the *HBB* gene with programmable proteins, variable outcomes have been reported. Targeting the LCR region of the β-globin locus

with the epigenetic regulator Cas9^{p300core} did not result in *HBB* gene activation in HEK293 cells [172], whereas targeted demethylation of specific CpGs (TALEN-TET1) in the promoter region strongly upregulated β-globin gene expression in K562 cells [171]. To the best of the researcher's knowledge, the effectiveness of programmable dCas9-VP16$_{(n)}$ for transcriptional activation of the β-globin gene has not yet been evaluated in iPSCs.

To date, different transfection approaches have been evaluated for delivery of Cas9 activators, including both transient non-integrative vector (dCas9-activator) expression [79, 95, 163, 164] or integrative approaches with stable or inducible dCas9-effector expression [166, 215, 216]. Although both systems were effective in achieving gene activation, transient transfection is usually used for achieving specific gene activation at a given DNA locus with limited number of guides, whereas stable cells lines are ideal for genome-scale screens with large gRNA libraries [165]. To screen for dCas9-gRNAs able to trigger specific target activation either singly or in multiplex, transient transfection with either liposomal (Lipofectamine) or non-liposomal (Fugene HD)-based transfection reagents in HEK293 cells is commonly used. However, the transfection conditions and amount of vector used relative to cell density vary widely between different studies.

Here, it was hypothesised that a dCas9-based approach would facilitate modelling of the RNA processing defect in cells with beta thalassaemia (IVSII-I G>A) and enable the assessment of the restoration of gene expression in iPSCs following gene correction. Thus, it was necessary to: (1) optimise a transfection protocol that would allow for screening of sgRNA activity in HEK293FT cells using pre-validated sgRNA combinations that target the promoter region of the *HBG* gene (2) assess if a similar strategy (synergistic effect of multiple guides) would enable transcriptional activation of the β-globin gene in HEK293FT cells, (3) evaluate the effectiveness of the strategy of *HBB* gene activation in iPSCs, and (4) compare *HBB* gene expression in beta thalassaemia iPSCs before and after gene correction.

Data indicated that at the cell density used in this study, the transfection of 1 µg of total vector (expressing gRNA-dCas9-activator) at a 3x charge ratio (Lipofectamine LTX) in HEK293FT yielded the maximum level of *HBG* gene induction (using a combination of pre-validated guides). The same strategy was applied to mediate gene activation of the *HBB* gene with either the dCas9VP64 or the VP160 system; however, increasing the TD tandem (VP16) repeat to a total of 10 repeats did not result in an increase in gene activation. This chapter also presents data that shows that this system can be applied to iPSCs, although in this cell type nucleofection enables a higher induction of gene expression than lipid-based transfection methods. Importantly, manipulating *HBB* expression with this system can model the splicing defect in beta thalassaemia iPSCs, as evidenced by the lack of normally spliced product upon gene activation in iPSCs with the beta thalassaemia

RNA processing mutation. Furthermore, this activation system can also be used to assess the restoration of the normal splicing process in gene-corrected iPSCs.

5.2 *Results and discussion*

5.2.1 Activation of globin genes (*HBG* and *HBB*) in HEK293FT cells using lipid-based transfection

5.2.1.1 *Gamma globin gene activation*

Several methods are available for transient transfection of cells with dCas9-expressing plasmids. They include chemical methods, like cationic liposomes [79], and non-liposome [164] or physical methods, such as electroporation [160]. However, cell viability and transfection efficiency varies with each transfection method [94, 217]. Efficient delivery is critical since in most cases robust dCas9-mediated gene activation requires the synergistic action of multiple guides. Indeed, single guides in conjunction with dCas9-VP16$_{(n)}$ systems are rarely associated with a considerable increase in gene expression [79, 164]. Transfection conditions reported in the literature varied widely (Table 5-1). It was therefore necessary to optimise the liposomal transfection conditions and screen for the guide combination that would achieve a sizeable *HBB* activation with dCas9-activator system in HEK293 cells. Four previously validated guide RNAs spanning the proximal promoter region of the *HBG1* and *HBG*2 genes were initially assessed, using Lipofectamine LTX-based transfection essentially as described by Perez-Pinera *et al.* [79].

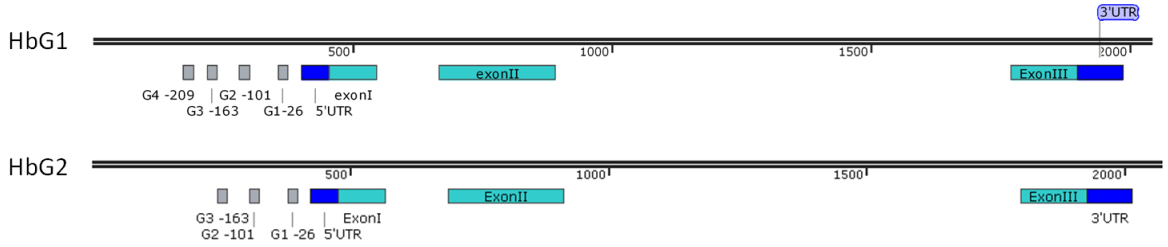

Figure 5-1: Schematic of the HBG1 and HBG2 genes and their proximal promoter regions. Grey boxes indicate the location of the targeted DNA sites, and the numbers indicate their position relative to the TSS. The guide sequences were reported in Perez-Pinera *et al.* [79].

Table 5-1: Transfection conditions used in some gene activation studies

Study	No. of transfected cells	Total plasmid	Charge ratio	Transfection reagent
Maeder *et al.*	$1.6x10^5$	0.5 µg	5x	Lipofectamine LTX
Konermann *et al.*	$0.2x10^5$	0.3 µg	2x	Lipofectamine 2000
Mali et al.	$4×10^5$	4 µg	NM	Lipofectamine 2000
Chavez *et al.*	$0.5x10^5$	0.210 µg	NM	Lipofectamine 2000
Hu et al.	$2.4x10^5$	1.6 µg	NM	Lipofectamine 2000

NM: not mentioned, charge ratio: transfection reagent/DNA ratio.

5.2.1.1.1 Activation of the *HBG* gene with sgRNA-dCas9-VP64-expressing vectors

Differences between this initial approach for lipid-based transfection and that published by Perez-Pinera *et al.* [79] are illustrated in Table 5-2. Principally, these researchers used two separate expression vectors; one for sgRNA, driven by the U6 promoter, and the other for dCas9-VP64, driven by the CMV promoter, with a mass ratio of 1:3, whereas in this study, the double-stranded oligonucleotide pairs of the gRNA were cloned into the BbsI site of a single vector expressing a sgRNA driven by the U6 promoter and a codon-optimised nuclease-deficient Cas9 fused to the VP64 acidic TD driven by the CBh promoter (addgene plasmid#48238) with a molar ratio of 1:1. The molar ratios of the expression plasmids in both studies are equivalent.

Table 5-2: Experimental conditions used in the current study compared to that reported by Perez-Pinera et al. [79]

	Perez-Pinera *et al.* [79]	**Current**
Cells	HEK293T	293FT
sgRNA	Four sgRNAs:G1, G2, G3, G4	
Format	NM	12 well
Lipofectamine(Lp)	2000	LTX/PLUS
Liposome/DNA ratio	NM	2.5:1
Total plasmid	NM	approximately 1 μg
No. of vectors used	Two (sgRNA and dCas9-VP64)	One (dual expression)
dCas9/gRNA	3:1 mass ratio (1:1 molar ratio)	1:1 molar ratio
q-PCR primer used	Span Intron 2	Span Intron 1

NM: not mentioned.

Although using a single vector with dual expression (sgRNA-dCas9VP64) has the advantage of ensuring the concomitant transfection of both components needed for gene activation, this approach does limit the manipulation of the levels of the individual elements. Consequently, three different experiments (A, B, and C) were initially carried out in order to screen for conditions that produced high gene expression levels using different guide combinations (Figure 5-2). The charge ratio of Lipofectamine to DNA ratio used was 2.5 to 1, as this permitted a high transfection efficiency as judged by GFP expression under fluorescence microscopy (Chapter IV, Section 4.2.1). In the first experiment (Experiment A) the total amount of plasmid DNA (pDNA) was kept constant at ≈1 μg. This amount was equally divided among multiple plasmids co-expressing dCas9VP64 and different sgRNAs, in order to study the effect of targeting multiple loci within the *HBG* promoter while keeping the total amount dCas9VP64 constant. In order to avoid wide variation between different conditions in terms of the total DNA amount or total gRNA, the transfection with the three and four guide combination was carried out at 1.2 μg total pDNA (≈300–330 ng per sgRNA-dCas9VP64 for 3 and 4 vectors multiplex vs 500 ng for two vectors). In Experiment (B) the amount of vector expressing individual sgRNAs (G1–G4) was kept constant (500 ng for each $sgRNA_{(n)}$-dCas9VP64)

as opposed to holding the total amount of pDNA constant and varying $sgRNA_{(n)}$ (Experiment A). The amount of the vectors used in Experiment (B) was replicated in Experiment (C) in the presence of a vector expressing a stronger activation domain (500ng dCas9-VP160, Addgene plasmid #48226). In the following sections $G_{(n)}$ will be used to denote $sgRNA_{(n)}$.

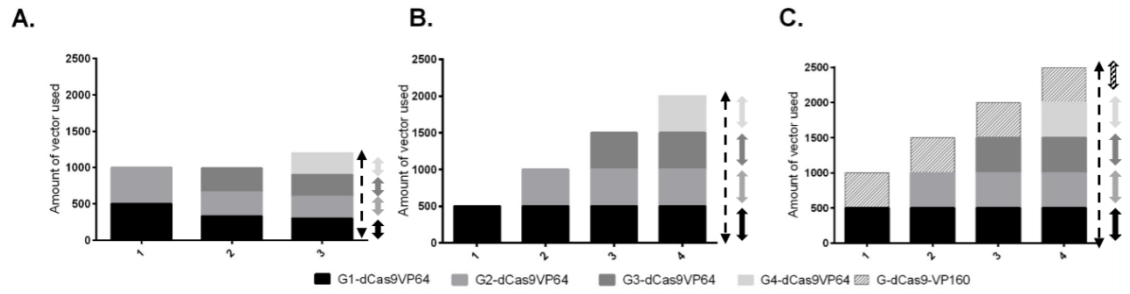

Figure 5-2: Schematic of the experimental design for gRNA-dCas9VP64-based HBG gene activation. Solid colours indicate the amount of individual gRNA-dCas9VP64 vector transfected relative to the total amount of the vectors used for transfections in Experiments A, B, and C. The total amount of dCas9VP64 is represented by the dashed arrows, while the amount of each guide is represented by the solid arrows.

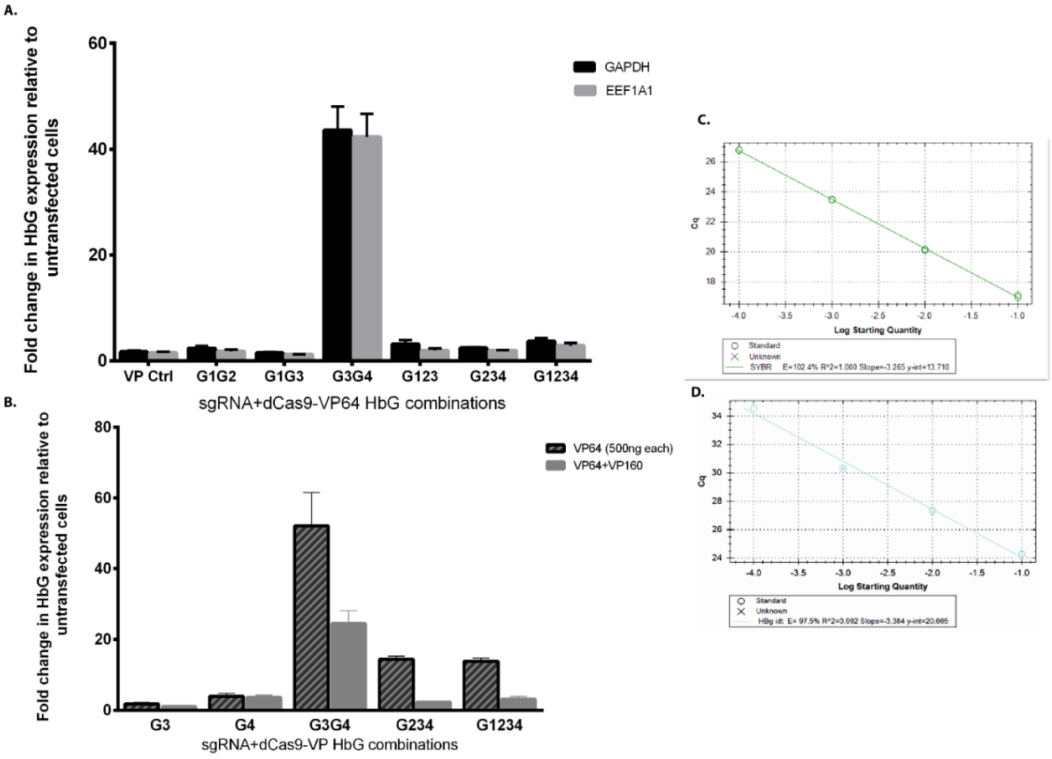

Figure 5-3: Activation of gamma globin gene expression in HEK293FT cells. The indicated vector or combination of sgRNA-dCas9VP64 vectors was used (A) with keeping a constant total of pDNA (~1 µg/well); (B) or by holding the amount of each sgRNA-dCas9VP64 constant (500 ng) in the presence (solid bars) or absence (pattern bars) of 500 ng of dCas9VP160. Primer efficiencies for GAPDH (C) and gamma globin (D) were 102.4 and 97.5 per cent, respectively. Data are shown as the mean for technical repeat±SEM.

Surprisingly, in Experiment (A), only a modest increase in gene expression (3.7-fold) was observed when the four sites were used to target the promoter region of the gamma globin genes (*HBG1* and *HBG2*), which is in contrast to the extent of activation (more than 100-fold change) reported by Perez-Pinera *et al.* [79]. A 40-fold increase was obtained upon targeting two sites at -163(G3) and -209(G4) from the TSS of the *HBG* genes (Figure 5-3.A). Similar data were obtained when different housekeeping genes (*EEF1A1/GAPDH*) were used for data normalisation. The primer efficiency was high, at 97.5 and 102.4 for *HBG* and *GAPDH*, respectively (Figure 5-3.C and D). Moreover, the levels of activation of gene expression between biological repeats was highly reproducible and displayed a consistent pattern, ruling out inconsistencies in transfection efficiencies (Figure 5-3.A and B). Indeed, the transfection efficiency of HEK293FT cells transfected with a GFP-expressing vector under similar conditions was consistently high, as assessed by fluorescence microscopy. The only possible difference between this study and that of Perez-Pinera *et al.* [79] is the cell line used; indeed, HEK293 clones are known to vary widely at both the genetic and epigenetic levels.

Using a single guide to target *HBG* did not result in substantial gene expression compared to multiplex guides (Figure 5-3.A and B), which is in agreement with previous studies on gRNA-dCas9-mediated gene activation [79, 95, 164]. Upon increasing the concentration of individual vectors expressing different sgRNA (sgRNA/dCas9-VP64) from 300 ng (Figure 5-3.A) to 500 ng (Figure 5-3.B) an increase in gene activation was observed, indicating a possible dose-dependent effect, which is consistent with other studies [163, 164]. The combination G2-4 increased gene expression from 2.4 (A) to 14.4 (B) and the combination G1-4 from 3.7 (A) to 13.9 (B) (Figure 5-3). However, it is uncertain whether this is due to the relatively higher levels of sgRNA, dCas9VP64, or both. Somewhat unexpectedly, a consistent decrease in gene expression was observed when the purportedly stronger activation domain (dCas9-VP160) was added to the combination with G34, G234, and G1234-dCas9VP64 (Figure 5-3.B). It is possible that this is due to increased toxicity of this plasmid, steric hindrance at the promoter, or DNA toxicity. Activation of the gamma globin gene was not due to generally increased transcription of the entire locus, since no mRNA expression for the beta or delta globin gene was detectable upon *HBG* gene activation.

5.2.1.1.2 Optimisation of the transfection conditions

(A) Amount of individual vector relative to total mass of plasmid

To further explore the previous observations, the guide combination (G34) that induced the highest level of increased *HBG* transcription was used to study the effect of varying the total amount of sgRNA and/or the total amount of pDNA on *HBG* expression levels. The level of gene activation was assessed using three different amounts of vector expressing individual sgRNAs (200 ng, 350 ng

and 500 ng) at three different levels of total pDNA (1 μg, 1.5 μg, and 2 μg), as illustrated in Figure 5-4.A. The manufacturer's recommendation for Lipofectamine LTX is between 1 μg and 1.5 μg for the plate format used in these experiments.

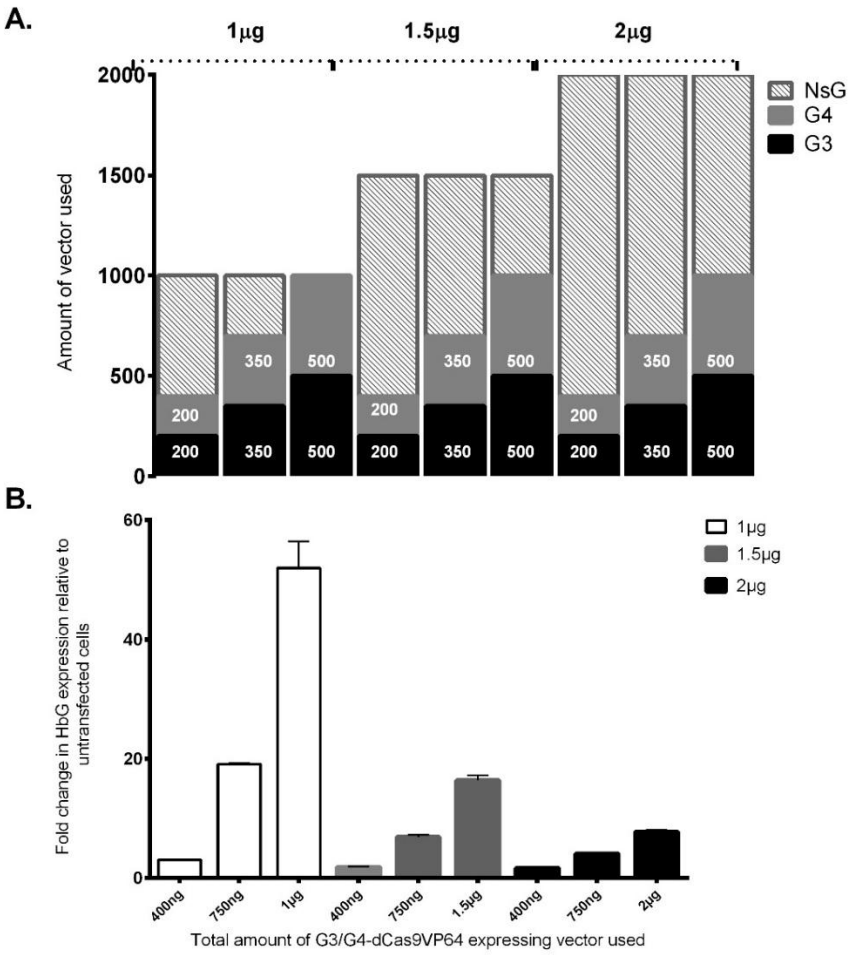

Figure 5-4: The effect of varying the amount of G34 and/or the total pDNA on HBG gene expression levels. (A) Schematic of the experimental design, where the solid bars indicate the mass of individual vectors expressing single sgRNAs transfected and patterned bars indicate the mass of empty vector NsG used to bring up the total to either 1 μg, 1.5 μg, or 2 μg, (B) RNA expression of *HBG* gene with the indicated amount of vector used. Data are shown as the mean for technical repeat±SEM.

The non-specific (non-binding) sgRNA-dCas9VP64 (NsG) vector or empty sgRNA was used to bring the amount of total pDNA to the desired constant value. In each of the conditions, the molar ratio of sgRNA relative to dCas9VP64 was kept identical. This experimental design facilitated the assessment of gene induction levels over a range of sgRNAs, while maintaining a constant amount of dCas9VP64 (constant total mass of plasmid). It also allowed the assessment of the effect of varying the total pDNA on the activity by holding the mass of vector expressing a specific guide constant. The experimental design should further reveal the maximum DNA amount tolerated by

the system, as DNA toxicity can be an issue with lipid-based transfection. The ratio of liposomal transfection reagent to pDNA was kept at 2.5 to 1.

The dose-dependent effect of sgRNA on gene expression activation is evident from the increase in expression within each subset of total pDNA used (Figure 5-4.B). This indicates that the apparent failure of combination G1234 to induce *HBG* gene expression as compared to combination G34 in Experiment (A) (Figure 5-3.A) was likely influenced by the reduced amount of sgRNA used; i.e., 300 ng vs 500 ng. Figure 5-4.B further shows that the activity induced by a similar mass of vectors expressing the most potent pair of gRNAs (G34) dropped sharply as the total amount of pDNA increased. For example, the change in *HBG* expression at 1 μg of G34 decreased by more than six-fold, from 51 to 7.7 when the total pDNA was increased from 1 μg to 2 μg. This observation may thus also explain the decrease in gene expression upon adding dCas9VP160 in Experiment (C) (Figure 5-3.B) at the specified charge ratio of liposome/DNA and indicates an apparent toxicity of higher plasmid loads, but does not rule out steric hindrance at the promoter.

(B) Charge ratio relative to total amount of plasmid

Selection of the charge ratio of 2.5:1 was based on GFP expression upon co-transfection with the GFP expression vector. To further investigate the potential effect of the charge ratio, gene expression levels were used as a read-out. *HBG* gene activation levels were analysed over a range of total amounts of pDNA while varying the charge ratio; i.e., the Lipofectamine to pDNA ratio (1x–5x). With this design, the increase in total pDNA is linked to a proportional increase in sgRNA and thus reveals the dose-dependent effect of the total pDNA (moles of pDNA=total active guide=dCas9VP64) (Figure 5-5.A). In other words, this approach should assess if the increase in the amount of expression vectors will be reflected by a parallel increase in gene expression across various ranges of charge ratios and total pDNA and thus validate the linear range of this system.

A statistically significant positive linear relationship between the amount of expression vector and the fold change in gene activation was observed upon using a charge ratio of 1x ($p=0.0051$, $R^2=0.976$) or 1.5x ($p=0.0016$, $R^2=0.98$) (Figure 5-5.C1 and 2) indicating that the increase in the amount of the expression vector is accompanied by a proportional increase in gene expression for these values of charge ratio over the entire range of pDNA under study. However, this relationship was not maintained over the range of 2x–5x. The limit of linearity for total pDNA decreased as the charge ratio increased (Figure 5-5.C). Therefore, increasing the total pDNA above 1 μg using 2.5x (as used in the previous experiments in this study) may affect gene expression due to suboptimal transfection conditions and therefore invalidate some of the comparisons presented in the previous sections. This explains to some extent the decrease in gene expression level observed upon adding

dCas9-VP160 in Experiment (C) and that observed for Experiments (A) and (B) with increasing total pDNA levels.

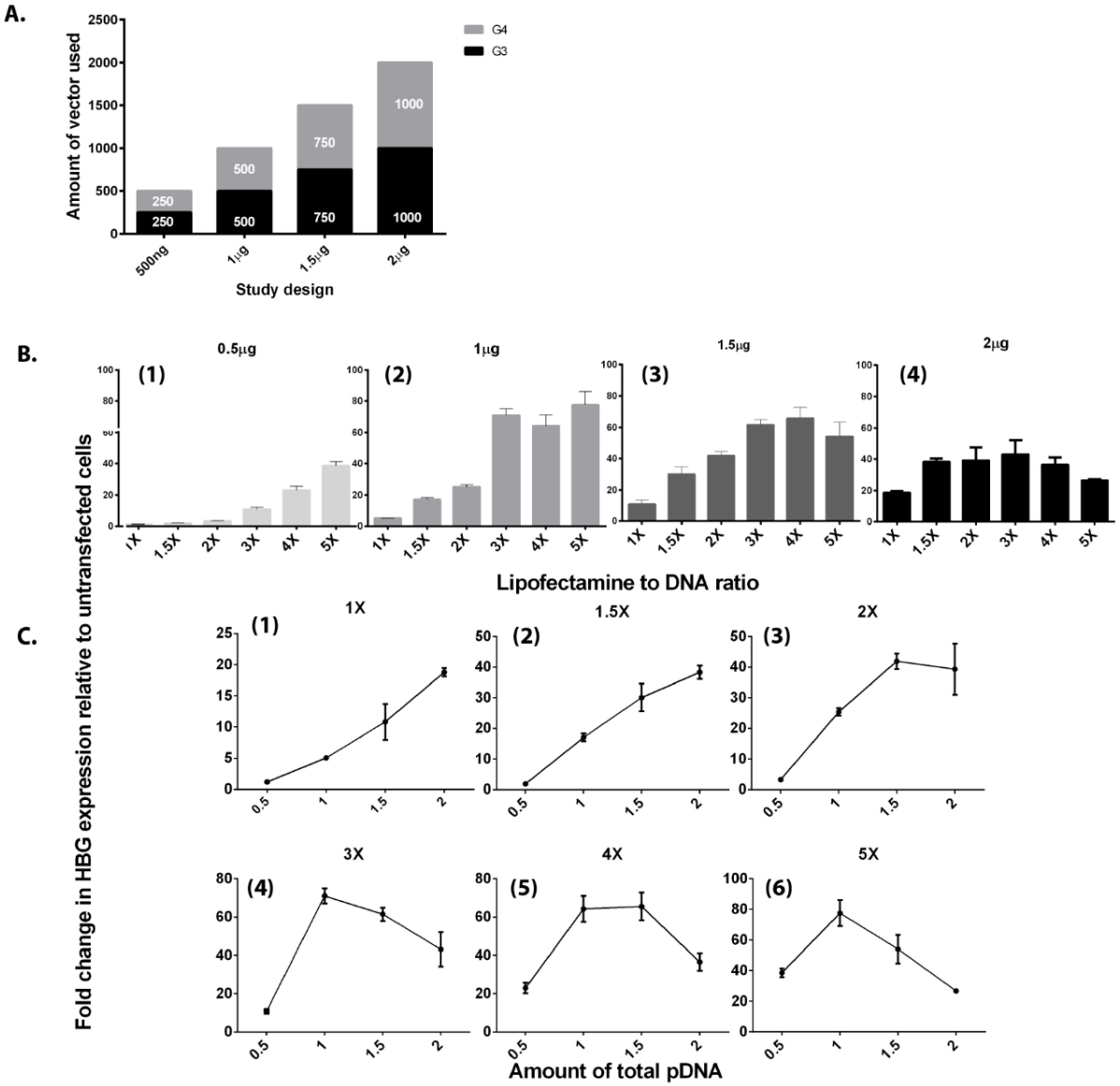

Figure 5-5: The effect of charge ratio (1x–5x) on HBG gene expression levels studied over a range of total pDNA (0.5–2 µg). (A) Schematic of the experimental design, where solid colours indicate the amount of transfected vector expressing the indicated sgRNA, (B) the effect of varying the charge ratio on RNA expression levels for cells transfected with the indicated total amount of pDNA, and (C) the effect of increasing the total amount of pDNA on RNA expression level at the indicated charge ratio. Data are shown as the mean for each technical repeat±SEM.

These data also indicate that the maximum efficiency of the system is observed at a Lipofectamine ratio of 3x using a total of 1 µg pDNA per well. A further increase in the expression vectors or change of the charge ratio did not result in an increase in gene expression.

Overall, these data indicate that at the cell density used in these experiments the maximum efficiency of a low range of total pDNA (500 ng) was obtained at a charge ratio of 5x. Under these

conditions linearity was maintained over the range of pDNA analysed (0.5–2 μg), and the maximum efficiency of the system, as assessed by the level of gene expression, was seen at 1 μg of total pDNA at 3x. For higher levels of plasmid DNA, it would be more appropriate to use a low charge ratio in order to keep the data in the linear range of the system (e.g., to eliminate possible toxicity). It is clear that these data provide a likely rationale for the variable Cas9-VP16$_{(n)}$ gene activation data reported in the literature. For example Maeder *et al.* [163] used a range of 500 ng to 1 μg of total plasmid at a charge ratio of 5x, which offers conditions that maintain the linear range while providing maximum efficiency of the system for the low range of pDNA. If one wishes to compare the effect of targeting multiple guides at the cell density and the well format used in this study, it is critical to maintain a constant total amount of pDNA at 1 μg with a charge ratio of 3x.

The initial experiment was repeated and the two and four guide combinations were compared using the conditions described above, while keeping the total plasmid amount at 1 μg (Figure 5-6.A). Another experimental condition was included, where cells were transfected with a total plasmid weight of 500 ng at a 5x charge ratio in the absence of an empty vector (NsG). Comparing the results obtained at different charge ratios will reflect the variability caused by experimental conditions for similar amount of guides. The results were consistent with those obtained for biological repeats under similar experimental conditions (1 μg/G34 [500 ng per guide] at 3x) (Figure 5-5.B2 and Figure 5-6.B). A dose-dependent effect of sgRNA on gene expression level was again observed for the two-vector combination G34, which is consistent with previous observations. However, in this case the activation level increased from 2.36 to 101.27 as the mass of individual vector (expressing G3 or G4) increased from 125 ng to 500 ng (Figure 5-6.B), which is in agreement with a previous study [163]. Interestingly, significantly higher activation of gene expression was observed upon using a charge ratio of 5x for a total pDNA of 500 ng for G34 compared to that obtained for similar amounts of vectors with a charge ratio of 3x for a total pDNA amount of 1 μg. This may indicate that including non-specific guide (50 per cent of total) may compete with the specific guide (25 per cent of total) to form the ribonucleoprotein. The data also suggests an inhibitory effect of G1 and G2 on the activity of G34, as the RNA expression level of *HBG* decreased by almost two-fold (8.7 to 4.8) under similar experimental conditions of charge ratio and total pDNA (Figure 5-6.B). Previous studies have demonstrated that gRNAs that target promoter regions close to TSS (mostly downstream) can interfere with RNA polymerase activity (CRISPRi) leading to partial or complete repression of activation [164, 167]. Although G1 for *HBG* is located upstream of the TSS, it overlaps with the TATA box, an RNA polymerase binding site, and thus may explain the observed repression of activation. How the sgRNA G2 (-101 upstream of

TSS) inhibits transcription induced by G3 and G4 remains unclear at present, but may involve interference with nucleosome positioning or Cas9-related characteristics. In fact, evidence suggests a long residency for Cas9 at the DNA binding site [71, 72] and also decreased potency in gene activation compared to transcription activator-like effectors (TALEs), which have been attributed to the intrinsic molecular characteristics of Cas9 [218].

The modest activation level observed with the four sgRNA combination is in agreement with that observed by Kabadi *et al.* [216], despite the fact that their system differs from the one used in these experiments (stable expression of dCasVP64 vs transient expression in this study).

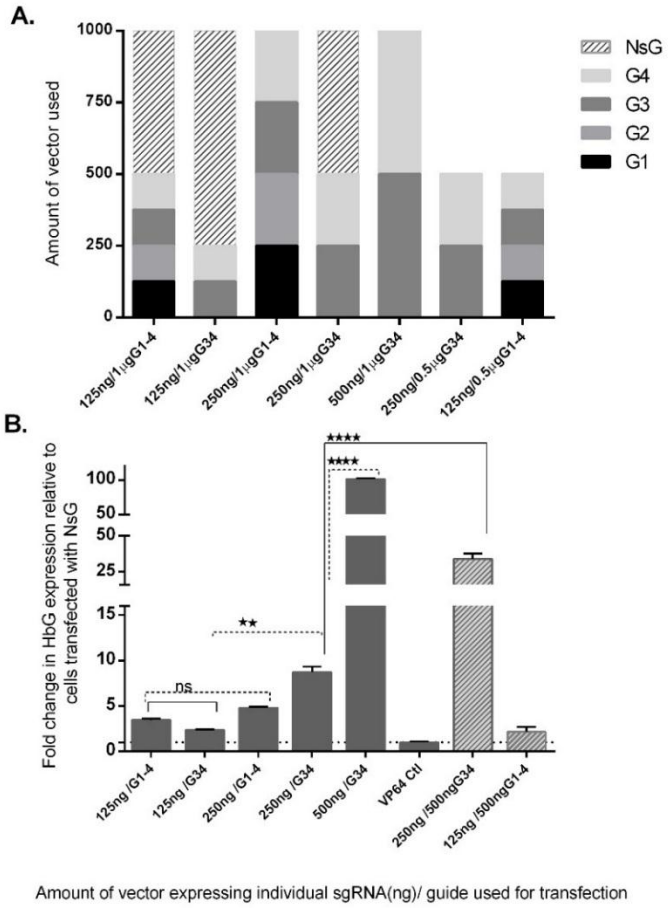

Figure 5-6: The RNA expression levels for the indicated guide combination carried out under similar experimental conditions. (A) Schematic of the experimental design, where solid colours indicate the amount of transfected vector expressing the indicated sgRNA. (B) RNA expression levels for the indicated amount of multiplex gRNA using a charge ratio of 3x and a constant amount of total pDNA. NsG: control vector, ns: nonsignificant. *p,0.05, **p<0.01 and ****p<0.001

The inconsistency in the level of gene expression induced by the four sgRNA combination between this study and that reported by Perez-Pinera *et al.* [79] clearly reflect the various levels of vector-related gene expression, charge ratios or variability in HEK293 cells lines. Moreover, the two guide

combination G34 was not examined in Perez-Pinera *et al.*'s [79] report. Holding the amount of total pDNA and sgRNA constant at an optimum pDNA/Lipofectamine ratio is needed for a valid comparison for synergistic experiments. Under the experimental conditions used in the present study, the maximum efficiency of induction was achieved using a total pDNA mass of 1 μg at a charge ratio of 3x for Lipofectamine LTX at this cell density. Lastly, different gRNA combinations targeting different site at the promoter region of *HBG* displayed variable levels of induction activity, which is consistent with previous studies of synergistic dCas9VP16$_{(n)}$-mediated gene activation [79, 163, 164, 216].

5.2.1.2 *β-globin gene activation*

5.2.1.2.1 Activation of the *HBB* gene with dCas9VP64 and dCas9VP160-expressing vectors

After optimising the conditions for inducing RNA expression for *HBG* gene in HEK293FT and validating the functionality of the system, the possibility of similar induction of the β-globin gene was explored.

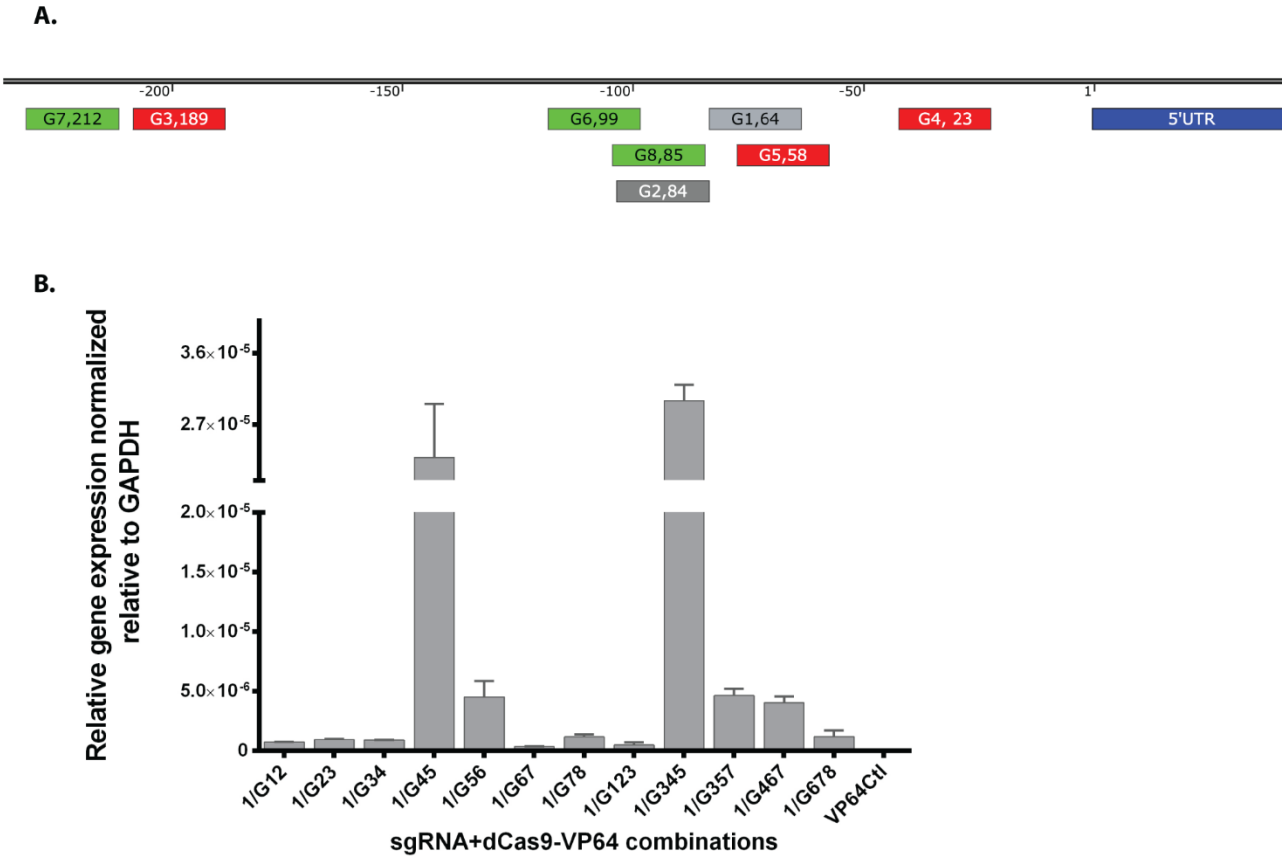

Figure 5-7: Activation of the HBB gene in HEK293FT cells. (A) Schematic of the targeting site for sgRNA in the proximal promoter region of the *HBB* gene. Red boxes indicate the sgRNAs with the highest activation efficiency in this set when tested in combination, while green boxes indicates those with a lower efficiency. Numbers indicate their position relative to the TSS. (B) RNA expression of the *HBB* gene using the VP64 system with the indicated gRNA.

To achieve this objective, it was first necessary to identify a combination sgRNAs capable of inducing expression of the β-globin gene at levels sufficient to allow the validation of expression of wild type *HBB* from the corrected gene in beta thalassaemia iPSCs.

Eight sgRNA were designed and cloned into the BbsI site of a single vector expressing both sgRNA and the dCas9VP64 vector. The guides span the proximal promoter region of the *HBB* gene (-23 to -212 from the TSS) (Figure 5-7.A). Combinations of two or three guides were transfected into HEK293FT cells in order to screen for their activity at inducing *HBB* gene expression (Figure 5-7.B). The total pDNA concentration was kept at 1 μg and was equally divided between the individual sgRNA used. The results indicated that Guides 3, 4 and 5 gave the highest gene induction efficiency among the combinations tested, with lower efficiencies observed for Guides 6, 7 and 8. Interestingly, using a guide (G4) located -20 bp from the TSS that overlaps the TATA box did not appear to repress the activity of Guides 3 and 5 (Figure 5-7.B), in contrast the repression seen for *HBG*.

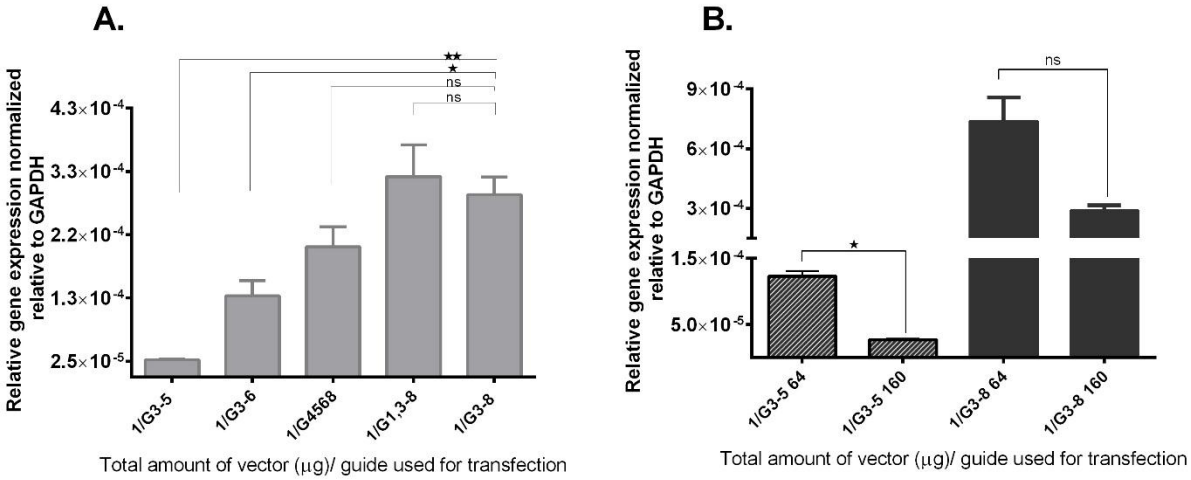

Figure 5-8: Activation of RNA expression from the HBB gene using the VP160 system. (A) Transcription activity associated with the indicated gRNA. (B) Comparison between the dCas9-VP64 or dCas9-VP160 systems for gene activation using the same sgRNA combinations. *p<0.05, and **p<0.01.

Having established that the sgRNA-VP16$_{(n)}$ system could induce *HBB* RNA expression, the next step was to assess whether a putative stronger activation domain could increase the activity at gRNA binding sites. Different permutations of sgRNA were selected and assessed for their activity using the dCas9VP160 system (Figure 5-8.A). The maximum value of gene activation with VP160 (under the tested conditions) was obtained for the combination G3-8. However, increasing the number of repeats of the VP16 activation domain from four to 10 again did not appear to have a substantial effect on the RNA expression level of the *HBB* gene, as the value for the same multiplex

of G345 was comparable ($3x10^{-5}$ for VP64 (Figure 5-7.B) and $2.7x10^{-5}$ for VP160 (Figure 5-8.A). To confirm this finding and rule out experimental variability, comparisons between both systems were carried out in parallel under similar transfection conditions, using two different combinations of sgRNA with a comparable mass of total plasmid (Figure 5-8.B). A slightly higher induction level was obtained with the VP64 system; however, this modest increase was not statistically significant for G3-8 (p=0.15) while the difference was marginally significant (p=0.047) for G3-5.

The statistical analysis of the data for off-target *HBG* gene expression levels in HEK293 cells (Figure 5-9.B) indicated that the increase in *HBG* gene expression was statistically significant (p=0.0003) for one combination (G3-8dCas9VP64) compared to untransfected controls. However, the same guide combination assessed with the dCas9VP160 system displayed no off-target activity. This discrepancy may be related to the potency of activation as displayed by the higher *HBB* induction level for G3-8-dCas9VP64. Overall, these results indicate that it is possible to induce *HBB* expression using $VP16_{(n)}$ systems, but that increasing the VP16 tandem repeat number in the activation domain does not lead to a significant difference in the induction of *HBB* expression for the guide combinations assessed under these experimental conditions.

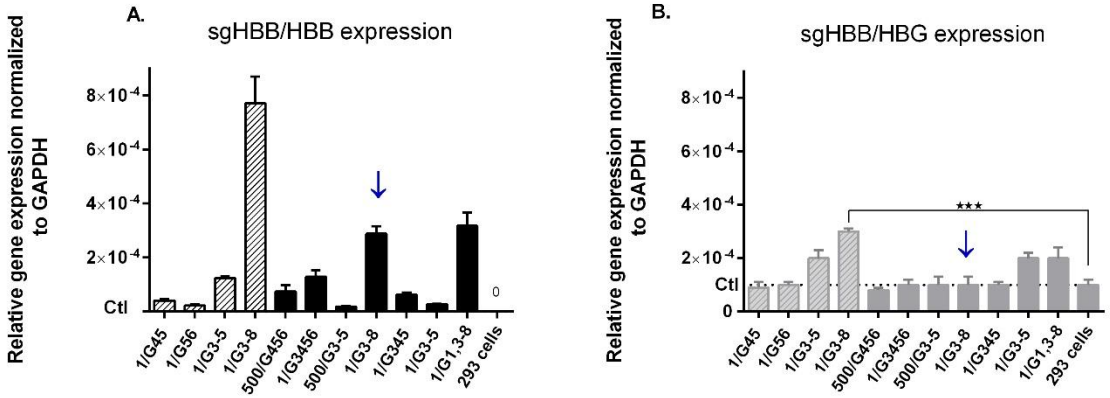

Figure 5-9: Off-target effects of different guide combinations targeting the HBB gene on the expression of HBG in HEK293FT. (A) Activation levels of the *HBB* gene by the indicated guide combination with either dCas9-VP64 (patterned bars) or dCas9-VP160 (solid bars) systems. (B) Non-specific activation of the *HBG* gene by the indicated guide targeting the promoter region of *HBB* with either dCas9-VP64 (patterned bars) or dCas9-VP160 (solid bars) systems. ***p<0.001

5.2.2 Activation of *HBB* in iPSCs

5.2.2.1 *Transfection of iPSCs with sgRNA(HBB)-dCas9-VP160 and off-target activity for the HBG gene*

To test whether the guides targeting the *HBB* promoter region are functional in iPSCs, two different active guide combinations (VP160 system) pre-validated in HEK293FT cells were assessed. A combination of vectors expressing all guides, with the exception of gRNA2 (G1, 3-8), or a combination that included gRNAs 4, 5, 6 and 8 were nucleofected into iPSCs. A plasmid carrying the green fluorescent protein (pmaxGFP, Lonza) was co-transfected with G4568 in order to assess the transfection efficiency. The latter was assessed using fluorescence microscopy (Figure 5-10.A). β-globin gene activation was measured 72 hours after transfection. The data indicate a significant activation of β-globin gene expression was induced in transfected iPSCs by the guided dCas9VP160 system (Figure 5-10.B). In this series of experiments in iPSCs the normalised relative gene expression levels were used instead of the fold change relative to untransfected cells (control) because no expression was detected in control cells. Transfection of dCas9VP160 with gRNAs did not affect *OCT4* or *NANOG* gene expression in the cell population. Some apparent off-target activation of the *HBG* gene was observed; however, the expression level did not reach statistical significance (p=0.442).

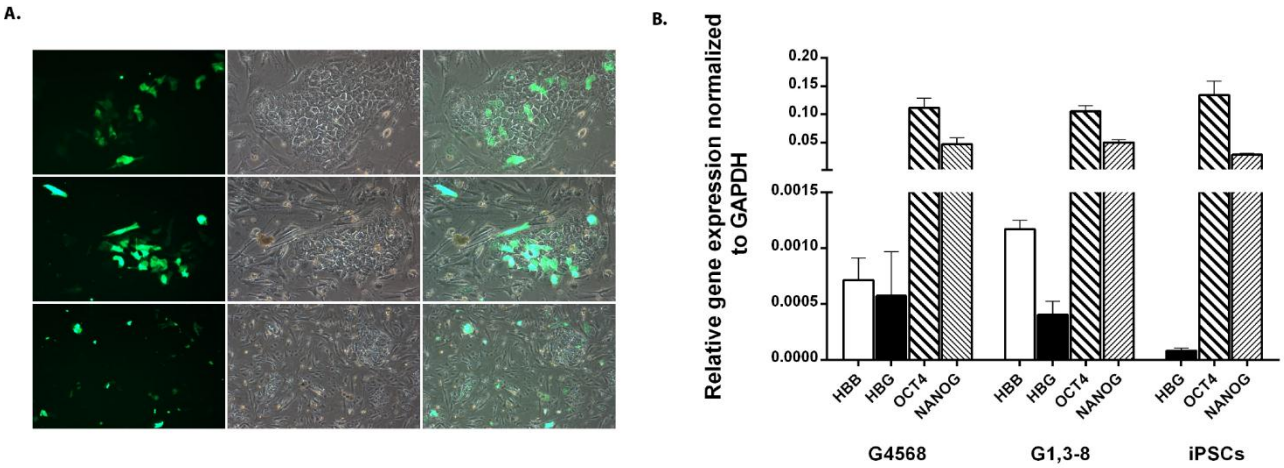

Figure 5-10: Activation of the HBB gene in iPSCs. (A) GFP fluorescence and phase contrast images of iPSCs nucleofected with G4568-dCas9VP160-expressing vectors and pmax GFP. (B) RNA expression levels of the *HBB*, *HBG*, *OCT4* and *NANOG* genes for iPSCs transfected with the indicated guide combination.

This small off-target activity was more pronounced when the four guide combination was used than with the seven guide combination, which may indicate a dose-dependent effect as the total pDNA was equally divided between the vectors expressing independent sgRNAs. Collectively, these data

suggested that the off-target effect of *HBB* sgRNAs on *HBG* expression may be attenuated by reducing the total amount of transfected vector in iPSCs and including the maximum number of tiled gRNAs. Therefore, subsequent experiments were carried with 1.2 μg of total vector (as opposed to 2 μg) using the VP160 system with a combination of vector that targets six sites in the promoter region (G3-8) and displays minimum non-specific activation of *HBG* (Figure 5-9, combination indicated by an arrow).

5.2.2.2 *Transfection method and HBB activation in gene-edited iPSCs compared to cells with beta thalassaemia*

Since the transfection efficiency of single cell iPSCs with nucleofection was low, transfecting iPSC colonies was tested in an effort to improve the efficiency of gene activation. Moreover, it was of interest to determine whether wild type *HBB* gene expression could be detected in corrected beta thalassaemia iPSC clones before exicising the selection cassette, a strategy that was deemed feasible because the selection cassette was inserted in the second intron region (Figure 5-11.A and B).

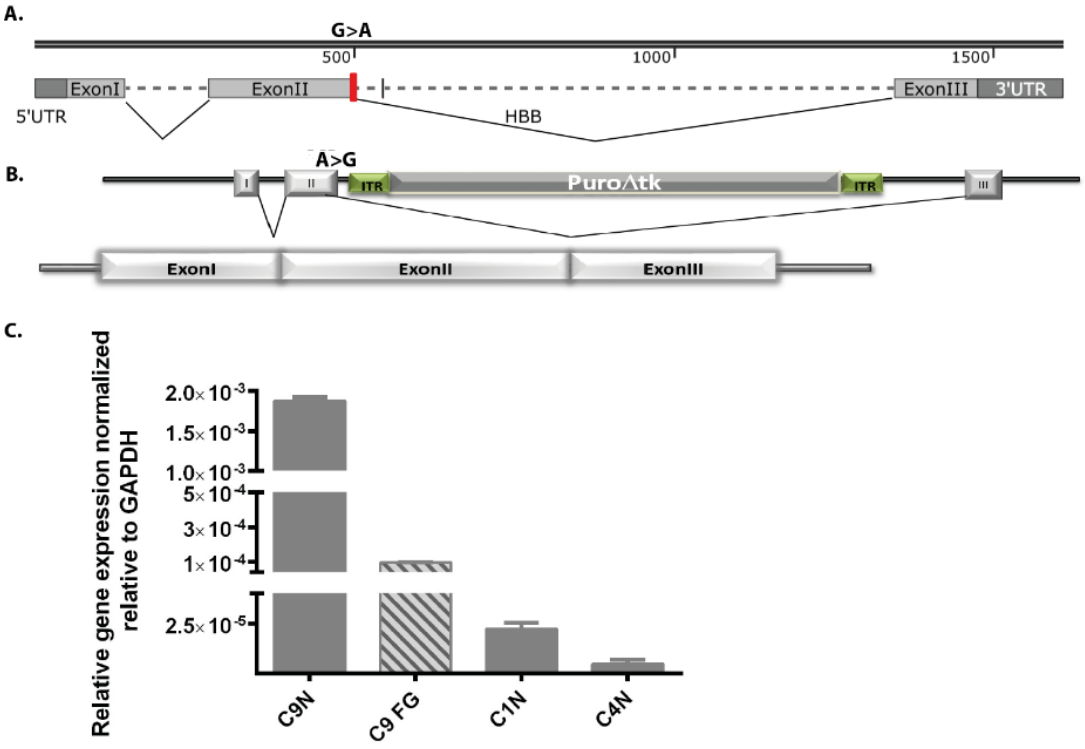

Figure 5-11: G(3-8)-dCas9VP160-mediated gene activation of the HBB gene in control iPSCs (wild type allele), iPSCs with beta thalassaemia and in biallelic gene-corrected iPSCs. (A) Schematic of the *HBB* gene showing the location of the RNA processing mutation (red box). (B) Schematic of the targeted *HBB* gene and the proposed splicing of the RNA transcript at the normal splice site, and (C) Relative gene expression levels obtained upon targeting the proximal promoter region with G(3-8)dCas9VP160 using either nucleofection (solid bars) or Fugene-based transfection (patterned bar). C9:control iPSCs, C1: iPSCs with beta thalassaemia, C4: iPSCs targeted at both alleles, FG: Fugene HD transfection, and N: nucleofection.

To this end, Fugene HD-based transfection or nucleofection was used to deliver 1.2 µg of the G(3-8)-dCas9VP160 vector combination in control iPSCs (wild type allele). Only nucleofection was used for comparing *HBB* gene activation levels in uncorrected iPSCs carrying the beta thalassaemia mutation and in gene-corrected beta thalassaemia iPSCs targeted at one or both β-globin alleles. The results indicated that nucleofection of colonies is a more efficient system for inducing gene expression compared to Fugene-based transfection. Surprisingly, a low level of gene activation in iPSCs with a splice junction mutation (beta thalassaemia mutation) was detected, albeit at a 100-fold lower level than in wild type control iPSCs (Figure 5-11.C). The melting profile of q-PCR products indicated the presence of a different product in iPSCs with beta thalassaemia compared to wild type cells (Figure 5-13.C). To further investigate the nature of this spliced transcript, a primer pair that spans intron II was used to PCR amplify the splice region from cDNA samples. The faint band observed for beta thalassaemia cells indicates an abnormal splice product with a size that is higher than the expected size of 79 bp. The size of the aberrantly spliced product was 126 bp indicating the use of an alternate splice site located downstream of the mutation site (Figure 5-12.B).

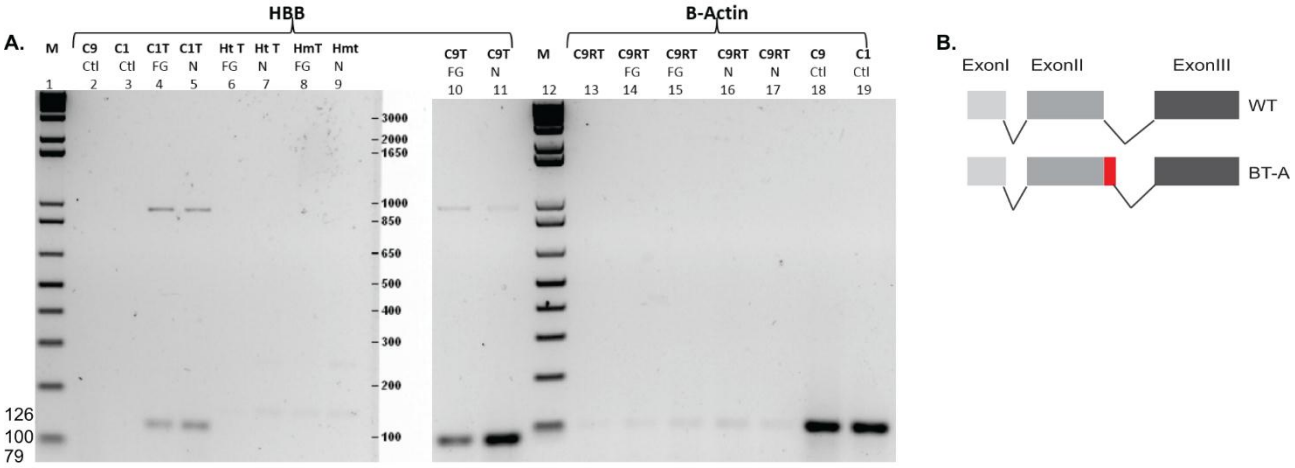

Figure 5-12: Aberrantly spliced products observed following RNA-guided gene activation of the HBB gene in iPSCs with beta thalassaemia or in gene-edited lines compared to control lines. (A) cDNA (RNA from cells transfected with G(3-8)-dCas9VP160) was used as a template for PCR. The size of the amplicons observed following gene activation in iPSCs with a beta thalassaemia mutation (126 bp and 929 bp Lanes 4 and 5), a monoallelic targeted line (Lanes 6 and 7), a biallelic targeted line (Lanes 8 and 9), a transfected control line (wild type allele, Lanes 10 and 11, 79bp), and the untransfected control iPSCs (Lanes 2 and 3). The presence of an extremely faint band in the 'no RT' control (RT, Lanes 13–17) for the indicated samples revealed an insignificant genomic DNA contamination compared to the robust band observed in cDNA samples (Lanes 18 and 19). (B) Schematic of the splicing process identified in wild type and mutant *HBB* gene, red box indicate the additional DNA fragment observed in the aberrant RNA product. C9:control iPSCs, C1: iPSCs with beta thalassaemia, Ht T: Transfected monoallelic, Hm T: transfected cells biallelic, RT: no reverse transcriptase, FG: Fugene HD transfection, N: nucleofection.

The larger band (929 bp) reflects an insignificant genomic DNA contamination. This is evident from the exceedingly faint band observed in the 'no reverse transcriptase' samples using primers in the exonic region of the housekeeping gene beta actin. Furthermore, although marginal gene expression levels were also observed in targeted cells, the size of the extremely faint band was also higher than that expected for a normally spliced product. Overall, these data indicate that the removal of the selection cassette from *HBB* gene-corrected iPSC is necessary, since its presence interferes with normal splicing of the transcript. More importantly, these data confirm that it is possible to model the splice junction defect in iPSCs with a beta thalassaemia mutation using the dCas9-activation system.

5.2.3 Gene expression in iPSCs after gene correction

To assess the restoration of normal splicing after gene correction and excision of the selection cassette, iPSCs derived from gene-edited iPSCs were transfected with the G(3-8)-dCas9VP160 vector combination and the gene expression levels were assessed using a primer pair that spans the intron II region of the *HBB* gene. Satisfyingly, the data consistently demonstrated the restoration of normal splicing (Figure 5-13.A). However, it was noted that the level of β-globin gene expression in gene-corrected clones remained significantly lower than in control cells. The similar levels of expression in two clones derived from the same parental iPSCs (BT-T iPSCs) targeted with two different guide combinations indicate a possible variability in the induction efficiency of gene expression between clones from different cell lines (control/beta thalassaemia) or different transfection efficiencies. Although confirmation of *HBB* expression at the protein level is needed, these observations indicate that gene-corrected beta thalassaemia iPSCs restored expression of *HBB* mRNA and that the presence of the introduced SNP near the splicing regions did not affect the splicing process. Moreover, reducing the total amount of the transfected plasmid was an effective strategy for mitigating the off-target effect observed in previous experiments. Indeed, the level of *HBG* gene expression in transfected cells was similar to that seen in untransfected cells (Figure 5-13.B). Melting curve analysis of the q-PCR products indicated the presence of a single peak for gene-corrected cells and controls (Figure 5-13.D). To further confirm and assess the type of the splice products of the *HBB* gene, PCR amplicons were generated from cDNA obtained from

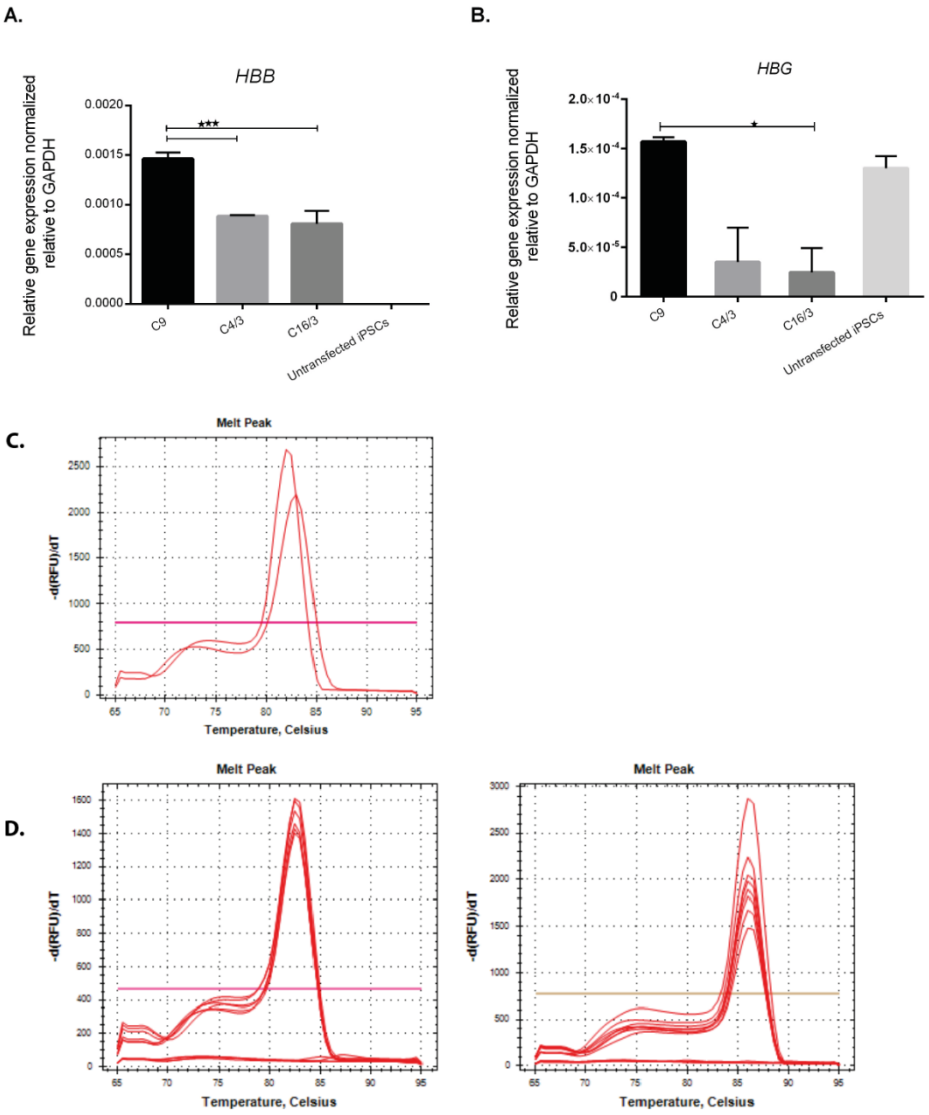

Figure 5-13: Restoration of normal splicing following the excision of the selection cassette and induced gene activation in transfected iPSCs compared to transfected parental beta thalassaemia iPSCs. (A) The levels of *HBB* gene expression in iPSCs determined 72 hours after transfection. (B) The level of off-target *HBG* gene expression. (C) Melt curves showing the different patterns for product amplified from iPSCs with beta thalassaemia compared to wild type. (D) Melt curve analysis showed overlaps of plots between gene-corrected iPSCs and wild type using two different primer pairs. C9: Control iPSCs, and C4/3 and C16/3: Gene-corrected beta thalassaemia iPSCs. Untransfected beta thalassaemia iPSCs. ***p<0.001

transfected gene-edited iPSCs. Similar band sizes were observed upon comparing the products obtained from transfected corrected iPSCs to those of the transfected control (the wild type allele). Thus, restoration of normal splicing in gene-corrected cell lines was achieved. Notably, the absence of bands in the no reverse transcriptase control indicated an absence of any genomic DNA contamination.

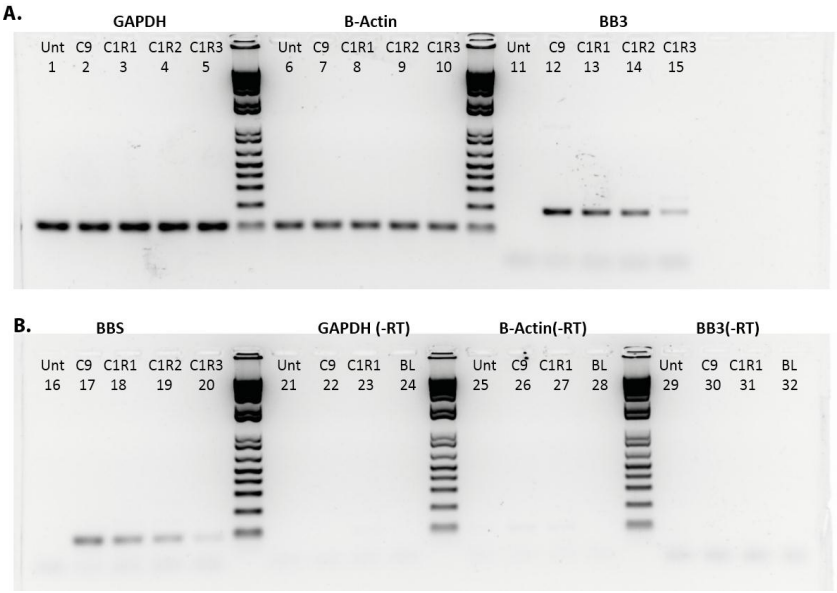

Figure 5-14: Normal splicing products observed following RNA-guided gene activation of the HBB gene in gene-corrected iPSCs. C9:Control iPSCs, C1: iPSCs with beta thalassaemia, C1R1, C1R2 and C1R3: Cells corrected at both alleles, BL: Blank PCR control, and –RT: no reverse transcriptase samples. Primer pairs used were exonic GAPDH (Lanes 1–5), exonic B-actin (Lanes 6–10), spanning intron II of the *HBB* gene: BB3 (Lanes 11–15), and BBS (Lanes 16–20).

Overall, these data suggest that no cryptic splice site was created by the SNPs and that the splicing occurred at the normal splicing site. Therefore, it can be concluded that the Cas9 activation-based strategy is effective for assessing the restoration of normal RNA processing in gene-corrected beta thalassaemia iPSCs.

5.3 *Conclusion*

This chapter describes a strategy for successful CRISPR-Cas9-mediated *HBB* gene activation in both HEK293FT cells and in iPSCs. The synergistic effect of using multiple gRNAs targeting cis regulatory elements has been previously demonstrated to be an effective approach to induce gene expression [79, 163, 164]. Building on these data, four pre-validated guide sequences that target the *HBG* gene [79] were selected and used to systematically optimise the transfection conditions of sgRNA-dCas9VP16$_{(n)}$ in HEK293FT cells. Interestingly, the data for *HBG* gene activation differed from those reported by Perez-Pinera *et al.* [79], who reported that the four combination guide resulted in a 100-fold increase, whereas in this study, a four-fold activation was achieved. This difference is not due to a limitation in VP64-mediated activation power since >100-fold gene activation was obtained by targeting the *HBG* promoter region with only two guides. The optimisation studies undertaken here have clearly demonstrated that it is critical in synergistic experiments to optimise factors that influence lipid-based transfection and that the total plasmid

load should ideally be kept constant using a neutral vector. The dose-dependent effects of sgRNA on gene activation levels observed in this study are in good agreement with those seen in previous studies [163, 164]. These findings are also in close agreement with a study by Kabadi *et al.* [216], who used an inducible genome-integrated version of the CRISPRa system [216]. It is likely that the difference between the current study and that of Perez-Pinera *et al.* [79] is related to such experimental details, or is perhaps due to intrinsic biological differences between the various HEK293FT clones used across the world.

These data indicate that the synergistic guide combination is also an effective approach to induce *HBB* globin gene expression using both the VP64 and VP160 systems. The purportedly stronger VP160 activation domain did not lead to a significant increase in gene expression over VP64 in this these experiments, an observation that concurs with a previous study [218]. The fact that activation of gene expression in corrected beta thalassaemia cells occurred at a lower level compared to that seen in control cells is somewhat puzzling. It is possible that this reduced efficiency is due to the zygosity of the locus, which can be readily verified by droplet digital PCR [219] or Southern blot, SNP location, variable transfection efficiency, or differences in passage number of the cells. This study has demonstrated that the CRISPRa system is certainly useful for modelling the RNA processing defect resulting from a base substitution at the donor splice site in beta thalassaemia iPSCs, and that it can be used to verify the absence of the normal spliced transcript in beta thalassaemia iPSCs. Importantly, the data from this study reveal that the splicing defect was restored in gene-corrected iPSCs, demonstrating the usefulness of this system for assessing the restoration of the normal splice site function of genes that are not necessarily expressed in undifferentiated iPSC. As such, this is a valuable tool to validate correctly targeted clones without the need for prolonged differentiation into cell types that express the gene of interest. However, it remains to be demonstrated that transient transfection of vectors in PSCs can also mediate an increase in protein expression, given the difficulty of achieving the efficient plasmid delivery needed for endogenous gene targeting in difficult-to-transfect human iPSCs and the variability in the induction efficiency. In fact, evidence suggests that an increase in gene expression may be associated with an increase in protein translation [79, 160, 166]. Although correlation between mRNA abundance and protein synthesis is variable [220], achieving a sizeable β-globin gene activation may translate into a higher level of protein expression.

CHAPTER VI

DISCUSSION

6 DISCUSSION AND FUTURE DIRECTIONS

The CRISPR/Cas9 system is a rapidly evolving technology that allows for site-specific genetic manipulation of iPSCs with remarkable flexibility and efficiency. Combining iPSCs and CRISPR/Cas9 platforms offers great potential for both basic and clinical research provided that efficient and precise genomic editing is achieved and the genomic integrity of cells is maintained. However, the genotoxicity of the nuclease system remains a major concern [55]. Studies that explored and validated proposed approaches to enhance both the accuracy of genome editing (i.e. decrease unwanted on-target modifications) and its specificity (i.e. decrease unwanted off-target modifications), in the context of gene correction in iPSCs, remain limited. Furthermore, CRISPRa has the potential to be used as a tool that can aid in modelling molecular defects, (e.g., those associated with RNA processing mutations) in addition to validating the restoration of normal transcription in gene-edited cells while circumventing the need for directed differentiation. In this study, it was hypothesized that it is possible to: (1) derive footprint-free iPSCs from dermal fibroblasts of a patients with homozygous mutations for beta thalassaemia using episomal-based reprogramming; (2) use a double nickase approach to avoid on-target modification, and a *piggyBac* transposon-based design for the donor template, to achieve a seamless selection cassette excision after the gene correction of the disease-causing mutation in iPSCs; and (3) use CRISPRa to activate the *HBB* gene allowing us to model the RNA processing defect, and to initiate normal transcription of the gene.

In this study, although transfection of dermal fibroblasts derived from three patients with homozygous IVSII-I (G>A) or IVSI-110 (G>A) mutations and a healthy donor was performed with two different combinations of episomal vectors (TPs: C4 or C19), the successful derivation of footprint-free iPSCs was achieved in only one patient with the beta thalassaemia mutation IVSII-I (G>A) and the control donor. The low efficiency of the episomal system observed in this study is consistent with previous reports [45, 46]. Episomes are lost by progressive dilution with successive cell divisions, as evidenced by the decrease in fluorescence from a GFP-expressing episomal vector to only 2.4% of cells at four weeks post-transfection (compared to 67.6% observed at week 1) [46]. The rapid loss of episomal vectors may limit the expression of reprogramming factors to the early stages of reprogramming, possibly contributing to the low efficiency of this approach. Even with further enhancement of the episomal-based approach by suppressing *TP53* and substituting *c-Myc* with *L-Myc*, the efficiency of the system remains significantly lower than other non-integrative approaches, such as the Sendai-virus or mRNA-based reprogramming [186], although one study has reported otherwise under feeder-free conditions [221]. The role of *TP53* suppression in promoting reprogramming has been documented in various studies; however, *TP53* plays an important role in

surveillance of the maintenance of genomic integrity [222]. Therefore, it is desirable to avoid using vector combinations with *TP53* suppression in order to allow it to carry out its quality control function during the process of reprogramming. Nevertheless, the oncogenic property of some of these factors (e.g, *cMyc* [223]) in the episomal plasmid combinations could be of concern. Under the experimental conditions of this study, the high attrition rate of putative iPSC clones was largely due to pseudo-iPSCs, abnormal karyotypes, complete reprogramming failure (in one sample with the IVSI-110 G>A mutation), or putative iPSC lines that failed to fulfil the stringent criteria of *bona fide* iPSCs. The failure of reprogramming in cells derived from one patient suggest that it may be due to sample-dependent factors, such as the establishment of the primary fibroblast cultures, since the reprogramming and cell cultures were performed in parallel with the other samples under similar conditions. Notably, and in agreement with a previous study [191], one vector combination yielded more clones under similar experimental conditions. However, the majority of the clones derived with this combination (C19) failed to show complete reprogramming. This may be related to variability associated with some technical aspects of transfection, such as the size and number of the transfected vectors. Alternatively, the timing, level and stoichiometry of the reprogramming factors have been reported to influence the success rate of reprogramming [45, 192, 224]. Both vector combinations carried the same number and type of transgenes; however, their arrangement varied, with possible subsequent consequences on their expression levels. Indeed, when enhanced transgene expression was achieved by including the woodchuck hepatitis virus post-transcriptional regulatory element in the design, an improved efficiency of reprogramming in human cord blood CD34+ cells was observed [225]. Moreover, these data corroborate evidence that highlights the importance of careful screening of iPSC clones for the presence of transgene integration or persistent expression of exogenous sequences [46, 186], as one of the iPSC clones (out of six screened clones) showed the integration of one of the reprogramming vectors.

High quality iPSCs are needed for realisation of the potential of iPSC-based cell therapies. However, evidence suggests that the combination of reprogramming factors used [226], the length of time in culture and the number of passages, in addition to the latency and the limited efficiency of the reprogramming process, are factors that may impact on the characteristics of iPSCs [18]. The highest efficiency for any of the non-integrative approaches was reported for mRNA-based reprogramming [186, 227]. This approximate the efficiency observed with the integrating viral vectors which have the highest efficiency of reprogramming; however, its advantage is negated by the requirement for repeated transfection [41] and the considerable failure rate in reprogramming [186]. Nevertheless, episomal vectors remain a viable option for reprogramming to pluripotency [186] that can be compatible with clinical translation [221].

Induced PSCs also represent an ideal platform for genetic engineering. Combining strategies that enhance the preciseness of available molecular biology tools (i.e. CRISPR/Cas9) can be useful in improving the on-target editing rate [145, 212] and in achieving highly specific genome engineering [100]. Precisely engineered iPSCs are needed for stem cell-based modelling of genetic disease, and for providing an unlimited source of gene-edited lines for potential use in regenerative medicine [228, 229]. Recent studies have reported successful gene correction of beta thalassaemia mutations using Cas9WT in iPSCs [35-37, 138]. Cooperative strategies, (i.e. double nickase (Cas9n) such as is used in the DN strategy), which induce DSBs at the on-target site while only single nicks are possible at each gRNA's off-target sites, would be more advantageous for precise DNA modification [96]. Because the on-target editing rate mediated by CRISPR/Cas9 is influenced by various factors that include the strategy design [145, 212], it was hypothesised that preventing the iterative cycles of dual Cas9n activity (introduction of DSBs) by either separation of the guide pair that targets the locus of interest [145] or by introducing a synonymous SNP in the PAM sequence [212] would be effective at limiting the on-target mutagenesis using the current approach. The double nickase strategy was explored using two different gRNA pairs in conjunction with a *piggyBac* transposon-based design for the donor vector. This study differed from that of Merkle *et al.* [145] in that the evaluation of the DN strategy with spatial separation of the gRNA pair was carried out in the context of achieving seamless gene correction with the additional limitations associated with the preference for using an endogenous, intronic TTAA sequence for insertion of the *piggyBac*-flanked selection cassette [182]. It also differed from Flynn *et al.* [212] in the donor construct design and the type of corrected mutation. The monogenic disease in the current study is autosomal recessive, whereas Flynn *et al.* [212] studied an X-linked disease (single copy, male iPSCs). This study demonstrates that the generation of bi-allelically targeted clones is achievable with this strategy. Interestingly, a higher percentage of targeted events was observed in the present study compared to a similar investigation, which used Cas9WT in conjunction with a *piggyBac*-based approach to correct a beta thalassaemia mutation [35]. Surprisingly, on-target mutations that resulted from a single nickase activity at the targeted allele were observed in some clones. Notwithstanding this observation, the data is still consistent with the notion that monomeric nickase activity is less prone to induction of indels compared to strategies that introduced DSBs. This notion is based on the observation of a high frequency of on-target mutagenic activity at the site of the induced composite DSBs (at the untargeted allele), whereas none were detected at the targeted allele at the site of the induced SSBs of one active guide (sgRNA3), while the frequency was considerably reduced at the other site (around 15 per cent at the sgRNA5 binding site). The absence of mutagenic activity at the binding site of the active nickase (sgRNA3) in one pair cannot be

attributed to inefficient activity, as evidenced by the high rate of mutagenesis at the untargeted allele in the presence of a closely placed nick (sgRNA1/sgRNA3). Although titrating the vector dosage has been reported to influence Cas9 activity [58, 75], the high frequency of on-target mutations observed at the untargeted allele could not be solely attributed to the amount of vector used in this study. Indeed, the data indicate that the expression of vector-borne genes is quite low, as evidenced by the weak fluorescent signal generated by the expression of EGFP fused to Cas9. This low expression can be attributed to a low transfection efficiency. However, the disproportionately high number of resistant clones obtained following selection, compared to the low number of GFP-positive cells observed, suggests that the low level of gene expression was sufficient to induce Cas9n activity. Notably, the strategy of antibiotic selection and clonal expansion can capture rare events that are possibly below the detection limit of methods commonly used to detect mutations in a heterogeneous cell population. Indeed, in this study, the frequency of on-target mutagenesis induced by DN (which produces composite DSBs) in the absence of donor was below the detection limit of the Surveyor assay in the bulk iPSC population. However, following clonal isolation of resistant clones, the majority of the untargeted alleles (heterozygous clones) carried an indel at the DN-targeted site.

Overall, the data in this study show: (1) a low gene editing efficiency in the bulk iPSC population; (2) that a correctly biallelic targeted clone can be achieved; (3) that the DN approach can afford high specificity; (4) high Cas9 DN-mediated on-target activity at the on-target site (untargeted allele); (5) that using a strategy that prevents on-target activity can improve the editing rate at the targeted allele; and (6) that although SSBs can be preferentially resolved by mutagenesis-free pathways, the features of the guide sequence or context can be a determining factor of mutagenic activity by Cas9n, albeit at a very low frequency compared to the full nuclease.

The difference in mutagenic activity of Cas9WT associated with a given guide sequence has been the subject of intense study [91, 210, 230, 231]. Various factors influence nuclease activity such as sgRNA characteristics, Cas9 protein characteristics, and the locus context of the target DNA sequence in the cell type under study. In fact, evidence suggests that the stability of sgRNA binding influences both on- and off-target Cas9 activities (as measured by indel induction efficiency) [76, 210]. RNA guide sequences rich in guanine bases displayed a higher efficiency than those rich in adenine bases [210]. Interestingly, the guide with the highest number of guanine bases (sgRNA5) was associated with mutagenesis. Alternatively, a long residency of Cas9 has been reported at the guide RNA-assisted DNA binding site [71, 72]. The intrinsic characteristics of Cas9, which may include its binding property, have been implicated in masking the guide binding site leading to a reduced potency of the dCas9-activator system at inducing transcriptional activation as compared to

the TALs systems [171]. Furthermore, the binding property of Cas9 has also been postulated to play a role in concealing the break site from detection by genome surveillance factors [71]. Based on the thermodynamic model of activity, decreasing the stability of binding by either changing the guide structure [92, 100], or by using a Cas9 variant with altered binding properties [69, 70], has been shown to decrease undesired mutagenic activity. A dimerisation-dependent approach (FokI-dCas9) retained a low level of activity as monomers [98]. Truncation of the guide by one nucleotide considerably reduced this undesired activity, while the removal by three nucleotides abolished the desired dimeric activity [100]. The proposed underlying mechanism for this observation was the inability to form a stable dimers. These strategies rely on the thermodynamic energy of binding to attenuate mismatch tolerance. One could, however, also argue the opposite in so much as given that both sgRNA and Cas9 can contribute to the stability of binding, it is possible that a guide sequences with increased stability can extend the Cas9 residency, leading to variability in the duration of the masking of the DNA break site, or can elicit other mechanisms of DNA repair to evict the bound protein with consequent variability in the resolution of the DNA break. A third possibility is that the context of the local sequence (e.g. its association with the nucleosome) of the targeted site can affect the rate and pathway of the repair by slowing the process, and thus shuttling the repair to resection-dependent pathways [131]. The concept of an interplay of factors not related to sequence properties that can alter the efficiency of the guide is best exemplified by the occasionally observed higher rate of mutagenesis at some off-target sites (with sequence mismatches) compared to the on-target site (sequence matches) at some loci [74].

This study has shown that the pattern of mutations generated with different guides with different offset varied significantly. Surprisingly, an inversion and insertion was observed in one clone targeted with a guide pair with an offset of 30 bp, whereas mostly deletions with micro-homology flanking the deleted sequence were associated with a guide pair with an offset of 5bp. The alignment of the sequence between the SSB sites in DNA using less than five base pairs of homology may indicate that the process occurred through C-NHEJ [110]. However, the presence of inversion and a relatively large insertion with the other guide pair indicate that the process occurred through the resection-dependent pathway. Although the number of the analysed clones, and targeted loci, were limited in this study, this observation may suggest that the guide pair design (e.g., offset length) can result in different mutational patterns, indicative of different mechanisms of DNA repair. Further, successful *piggyBac*-mediated biallelic excision of the selection cassette were seen, with full and traceless *PiggyBac* transposon removal resulting in reconstitution of the corrected *HBB* allele. Additionally, no detectable off-target effects for nickase was detected upon screening the top predicted sites based on sequence homology.

Overall, these data on gene correction indicate that traceless gene editing is achievable at both alleles in iPSCs using a precise editing approach. Both strategies of DN activity on the corrected allele used in this study have proven effective in reducing unwanted on-target activity. Precise genome editing will harness the potential of nuclease-induced stimulation of repair mechanisms by allowing accurate site-specific editing of the genetic mutation while retaining the context of its endogenous genomic elements. The ideal cell-based therapy approach will have to preserve genome integrity. The data from this study, and those of others [145], show that even single nickase (SpCas9N) can be associated with low level mutagenic activity in iPSCs, implying that similar activity can occur at off-target sites. We examined the top predicted off-target sites for Cas9n activity, but recognise that for future studies undertaking a comprehensive analysis of off-target modifications (such as WGS) is a highly desirable validation for this approach. For genome edited cells to become clinically relevant the entire gene correction process would need to be carried out using processes that are compatible with good manufacturing practice (cGMP) that include defined media and xenogenic- and feeder-free processes which, at present, are a challenge. The combination of strategies used in this study can be further enhanced by using the truncated RNA guides [92], reducing the amount of vectors [75], and by using culture conditions that improve tolerance to single cell manipulation [232].

Disease modelling of beta haemoglobinopathies remains unattained as erythroid cells obtained via direct differentiation of iPSCs retain an immature phenotype with little expression of the β-globin gene [148-152, 154]. This study sought to assess the applicability of CRISPRa for stimulating gene expression of the β-globin locus inactive in iPSCs, explore the potential of this system for modelling the RNA processing defect in beta thalassaemia and validate the restoration of functionality of the wild type allele in gene-corrected iPSCs. The results shows that it is possible to use the CRISPR/Cas9-assisted genome editing approach to demonstrate restoration of appropriate gene expression after gene correction, and its absence in iPSCs with the disease mutation. This indicates that a similar approach can be used to model other diseases, including those caused by mutations that interfere with normal RNA processing. Confirmation of the zygosity of the locus is necessary to determine if the observed reduced expression of the normal transcript reflects suboptimal locus activation as a result of the transfection efficiency or is a consequence of differences in epigenetic effects at the *HBB* locus. Western blotting will also be employed to assess β-globin protein translation in CRISPRa-activated iPSCs.

In conclusion, this study has optimized and validated an approach that can be used to achieve traceless gene correction with enhanced specificity and presents a strategy that permits assessment of the success of the gene correction in iPSCs (in this case *HBB*) without the need to differentiate

the iPSCs into specific types of cells (while recognizing that it is still necessary to demonstrate β-globin expression at the protein level in terminally-differentiated cells). This approach may provide an example of a viable platform for site-specific gene correction for the locus silenced in hPSCs, loci with a low efficiency of editing rate as the rate varied over a broad range at different loci [145]. Achieving a high efficiency of targeting can obviate the need for positive/negative selection, and allow the identification of corrected clones by direct analysis [233]. The suggestion to exploit the high efficiency of NHEJ in achieving a beneficial loss of function of some DNA elements [8, 234] can offer a viable approach to overcome the limitation posed by the low HDR editing rate observed in quiescent autologous-HSCs [5, 234]. For example, inactivation of a lineage specific enhancer of BCL11A can result in an induction of HbF synthesis [235] to such a level that it has the potential to ameliorate the phenotype associated with beta haemoglobinopathies [236].

The presence of unwanted mutations in stem cell models of genetic disease can confound the association of a certain genotype with its observed phenotype [55]. It can also complicate the generation of reliable isogenic controls which are needed for the investigation of the molecular mechanisms underlying the pathogenesis of the disease [202]. More importantly, in clinical settings even a low, undetectable mutation level can have undesired consequences. This study has demonstrated that even the nickase form of SpCas9 can be associated with a low level of mutagenesis in iPSCs. Therefore, a DN strategy still carries the risk of off-target activity that can be associated with monomeric nickase, albeit at a significantly lower level compared to Cas9WT [96]. The dimerization-dependent approach (*Fok*I-dCas9) has also been demonstrated to have monomeric activity, that can be the result of a reaction mechanism of *Fok*I [237]. Remarkably, *Fok*I activity was shown not to be restricted by spacer size as they can also dimerize across a significant distance [237]. Nevertheless, the average frequency of indel associated with monomeric *Fok*I-dCas9 activity was less than that observed with monomeric nickase [98].

The rapid pace of research in this field has recently provided additional tools that can further enhance the specificity of inducing DNA breaks to stimulate DNA repair mechanisms. A new DNA-guided endonuclease identified in *N. gregoryi* SP2 (Natronobacterium gregoryi SP2), denoted ‘NgAgo’, was demonstrated to induce endonuclease activity that was considerably decreased in the presence of even a single mismatch [238]. It was also demonstrated to be absolutely intolerant to a three consecutive base mismatch, which is a major advantage over the Cas9 system, as it significantly limits the screen of the intactness of off-target sites. The system is not restricted by the presence of an adjacent motif, and possibly has a higher tolerance of high GC regions as judged from the observed activity at the *HBA2* locus. More importantly, the size of NgAgo is smaller than that of the SpCas9 system, which can be more amenable to packaging in vectors for *in vivo*

delivery. Although recently a small size Cas9 orthologue has been proposed to be useful for *in vivo* gene editing [239], Cas9 derived from *Staphylococcus aureus* (SaCas9) is restricted by the requirement for the PAM sequence needed for activity. NgAgo could be a promising tool, provided that aspects such as the transient time, iterative cycle of activity, off-target activity and the versatility of this tool prove advantageous for genome engineering. The rapid progress in this field has provided a wide array of tools and approaches for genome engineering that may facilitate moving this field towards therapeutic applications. Although various bioinformatic tools are available that rank guide sequences based on predicted scores for on-/off-target specificity [54], the accuracy or specificity of their performance can be affected by factors not yet completely understood. We still lack comprehensive methods that will allow us to reliably assess a low level of mutation [55], and an effective *in silico* and *in vitro* approach that can reduce the empirical testing and validation of the off-target activity predicted in the cellular context.

7 REFERENCES

1. Modell, B. and M. Darlison. (2008). Global epidemiology of haemoglobin disorders and derived service indicators. Bull World Health Organ, 86(6): p. 480-7.
2. Michlitsch, J.G. and M.C. Walters. (2008). Recent advances in bone marrow transplantation in hemoglobinopathies. Curr Mol Med, 8(7): p. 675-89.
3. Finotti, A., et al. (2015). Recent trends in the gene therapy of beta-thalassemia. J Blood Med, 6: p. 69-85.
4. Naldini, L. (2015). Gene therapy returns to centre stage. Nature, 526(7573): p. 351-60.
5. Genovese, P., et al. (2014). Targeted genome editing in human repopulating haematopoietic stem cells. Nature, 510(7504): p. 235-40.
6. Hoban, M.D., et al. (2015). Correction of the sickle cell disease mutation in human hematopoietic stem/progenitor cells. Blood, 125(17): p. 2597-604.
7. Chapman, J.R., M.R. Taylor, and S.J. Boulton. (2012). Playing the end game: DNA double-strand break repair pathway choice. Mol Cell, 47(4): p. 497-510.
8. Cox, D.B., R.J. Platt, and F. Zhang. (2015). Therapeutic genome editing: prospects and challenges. Nat Med, 21(2): p. 121-31.
9. Cao, A. and R. Galanello. (2010). Beta-thalassemia. Genet Med, 12(2): p. 61-76.
10. Glaser, A., B. McColl, and J. Vadolas. (2015). The therapeutic potential of genome editing for beta-thalassemia. F1000Res, 4.
11. Bauer, D.E. and S.H. Orkin. (2015). Hemoglobin switching's surprise: the versatile transcription factor BCL11A is a master repressor of fetal hemoglobin. Curr Opin Genet Dev, 33: p. 62-70.
12. Mansilla-Soto, J., et al. (2016). Cell and Gene Therapy for the Beta-Thalassemias: Advances and Prospects. Hum Gene Ther, 27(4): p. 295-304.
13. Galanello, R. and R. Origa. (2010). Beta-thalassemia. Orphanet J Rare Dis, 5: p. 11.
14. Thein, S.L. (2013). The molecular basis of beta-thalassemia. Cold Spring Harb Perspect Med, 3(5): p. a011700.
15. Mettananda, S., R.J. Gibbons, and D.R. Higgs. (2015). alpha-Globin as a molecular target in the treatment of beta-thalassemia. Blood, 125(24): p. 3694-701.
16. Takahashi, K., et al. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell, 131(5): p. 861-72.
17. Yu, J., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science, 318.
18. Alateeq, S., P.R. Fortuna, and E. Wolvetang. (2015). Advances in reprogramming to pluripotency. Curr Stem Cell Res Ther, 10(3): p. 193-207.
19. Kelley, J.M. and G.Q. Daley. (2013). Hematopoietic defects and iPSC disease modeling: lessons learned. Immunol Lett, 155(1-2): p. 18-20.
20. Choi, K.D., et al. (2009). Hematopoietic and endothelial differentiation of human induced pluripotent stem cells. Stem Cells, 27(3): p. 559-67.
21. Vodyanik, M.A. and Slukvin, II. (2007). Hematoendothelial differentiation of human embryonic stem cells. Curr Protoc Cell Biol, Chapter 23: p. Unit 23 6.
22. Panopoulos, A.D. and J.C. Belmonte. (2012). Induced pluripotent stem cells in clinical hematology: potentials, progress, and remaining obstacles. Curr Opin Hematol, 19(4): p. 256-60.
23. Hanna, J., et al. (2007). Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. Science, 318(5858): p. 1920-3.
24. Vo, L.T. and G.Q. Daley. (2015). De novo generation of HSCs from somatic and pluripotent stem cell sources. Blood, 125(17): p. 2641-2648.

25. Riviere, I., C.E. Dunbar, and M. Sadelain. (2012). Hematopoietic stem cell engineering at a crossroads. Blood, 119(5): p. 1107-16.
26. Wang, Y., et al. (2012). Genetic correction of beta-thalassemia patient-specific iPS cells and its use in improving hemoglobin production in irradiated SCID mice. Cell Res, 22(4): p. 637-48.
27. Papapetrou, E.P., et al. (2011). Genomic safe harbors permit high beta-globin transgene expression in thalassemia induced pluripotent stem cells. Nat Biotechnol, 29(1): p. 73-8.
28. Ye, L., et al. (2009). Induced pluripotent stem cells offer new approach to therapy in thalassemia and sickle cell anemia and option in prenatal diagnosis in genetic diseases. Proc Natl Acad Sci U S A, 106(24): p. 9826-9830.
29. Varas, F., et al. (2009). Fibroblast-derived induced pluripotent stem cells show no common retroviral vector insertions. Stem Cells, 27(2): p. 300-6.
30. Sommer, C.A., et al. (2012). Residual expression of reprogramming factors affects the transcriptional program and epigenetic signatures of induced pluripotent stem cells. PLoS One, 7(12): p. e51711.
31. Koyanagi-Aoi, M., et al. (2013). Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc Natl Acad Sci U S A, 110(51): p. 20569-74.
32. Zhao, T., et al. (2011). Immunogenicity of induced pluripotent stem cells. Nature, 474(7350): p. 212-5.
33. Morizane, A., et al. (2013). Direct Comparison of Autologous and Allogeneic Transplantation of iPSC-Derived Neural Cells in the Brain of a Nonhuman Primate. Stem Cell Reports, 1(4): p. 283-292.
34. Gonzalez, F., S. Boue, and J.C. Izpisua Belmonte. (2011). Methods for making induced pluripotent stem cells: reprogramming a la carte. Nat Rev Genet, 12(4): p. 231-42.
35. Xie, F., et al. (2014). Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. Genome research.
36. Song, B., et al. (2015). Improved hematopoietic differentiation efficiency of gene-corrected beta-thalassemia induced pluripotent stem cells by CRISPR/Cas9 system. Stem Cells Dev, 24(9): p. 1053-65.
37. Yang, Y., et al. (2016). Naïve Induced Pluripotent Stem Cells Generated From β-Thalassemia Fibroblasts Allow Efficient Gene Correction With CRISPR/Cas9. Stem Cells Translational Medicine, 5(1): p. 8-19.
38. Ma, N., et al. (2013). Transcription activator-like effector nuclease (TALEN)-mediated gene correction in integration-free beta-thalassemia induced pluripotent stem cells. J Biol Chem, 288(48): p. 34671-9.
39. Sommer, C.A., et al. (2009). Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. Stem Cells, 27(3): p. 543-9.
40. Meier, I.D., et al. (2010). Short DNA sequences inserted for gene targeting can accidentally interfere with off-target gene expression. FASEB J, 24(6): p. 1714-24.
41. Mandal, P.K. and D.J. Rossi. (2013). Reprogramming human fibroblasts to pluripotency using modified mRNA. Nature Protocols, 8(3): p. 568-582.
42. Kim, D., et al. (2009). Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. Cell Stem Cell, 4(6): p. 472-6.
43. Cho, H.J., et al. (2010). Induction of pluripotent stem cells from adult somatic cells by protein-based reprogramming without genetic manipulation. Blood, 116(3): p. 386-95.
44. Ren, C., et al. (2006). Establishment and applications of epstein-barr virus-based episomal vectors in human embryonic stem cells. Stem Cells, 24(5): p. 1338-47.
45. Yu, J., et al. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science, 324(5928): p. 797-801.

46. Okita, K., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat Methods, 8(5): p. 409-12.
47. Hu, K., et al. (2011). Efficient generation of transgene-free induced pluripotent stem cells from normal and neoplastic bone marrow and cord blood mononuclear cells. Blood, 117(14): p. e109-19.
48. Dowey, S.N., et al. (2012). Generation of integration-free human induced pluripotent stem cells from postnatal blood mononuclear cells by plasmid vector expression. Nat Protoc, 7(11): p. 2013-21.
49. Frappier, L. (2012). The Epstein-Barr Virus EBNA1 Protein. Scientifica (Cairo), 2012: p. 438204.
50. Singbrant, S., et al. (2015). Two new routes to make blood: Hematopoietic specification from pluripotent cell lines versus reprogramming of somatic cells. Exp Hematol, 43(9): p. 756-9.
51. Hockemeyer, D. and R. Jaenisch. (2010). Gene targeting in human pluripotent cells. Cold Spring Harb Symp Quant Biol, 75: p. 201-9.
52. Kim, H. and J.S. Kim. (2014). A guide to genome engineering with programmable nucleases. Nat Rev Genet, 15(5): p. 321-34.
53. Gaj, T., C.A. Gersbach, and C.F. Barbas, 3rd. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol, 31(7): p. 397-405.
54. Lee, C.M., et al. (2016). Nuclease Target Site Selection for Maximizing On-target Activity and Minimizing Off-target Effects in Genome Editing. Mol Ther, 24(3): p. 475-87.
55. Tsai, S.Q. and J.K. Joung. (2016). Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. Nat Rev Genet, 17(5): p. 300-12.
56. Sander, J.D. and J.K. Joung. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol, 32(4): p. 347-55.
57. Silva, G., et al. (2011). Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. Curr Gene Ther, 11(1): p. 11-27.
58. Hsu, P.D., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol, 31(9): p. 827-32.
59. Jinek, M., et al. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science, 337(6096): p. 816-21.
60. Jinek, M., et al. (2013). RNA-programmed genome editing in human cells. Elife, 2: p. e00471.
61. Cho, S.W., et al. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res, 24(1): p. 132-41.
62. Mali, P., et al. (2013). RNA-guided human genome engineering via Cas9. Science, 339(6121): p. 823-6.
63. Cong, L., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science, 339(6121): p. 819-23.
64. Cho, S.W., et al. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol, 31(3): p. 230-2.
65. Tsai, S.Q., et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol, 33(2): p. 187-97.
66. Zhang, Y., et al. (2014). Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. Sci Rep, 4: p. 5405.
67. Wright, A.V., et al. (2015). Rational design of a split-Cas9 enzyme complex. Proc Natl Acad Sci U S A, 112(10): p. 2984-9.
68. Nishimasu, H., et al. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell, 156(5): p. 935-49.

69. Slaymaker, I.M., et al. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science, 351(6268): p. 84-8.
70. Kleinstiver, B.P., et al. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature, 529(7587): p. 490-5.
71. Richardson, C.D., et al. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat Biotechnol, 34(3): p. 339-44.
72. Sternberg, S.H., et al. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature, 507(7490): p. 62-7.
73. Shui, B., et al. (2016). The Rise of CRISPR/Cas for Genome Editing in Stem Cells. Stem Cells Int, 2016: p. 8140168.
74. Fu, Y., et al. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol, 31(9): p. 822-6.
75. Pattanayak, V., et al. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol, 31(9): p. 839-43.
76. Lin, Y., et al. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res, 42(11): p. 7473-85.
77. Li, H.L., et al. (2015). Precise correction of the dystrophin gene in duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPR-Cas9. Stem Cell Reports, 4(1): p. 143-54.
78. Yang, L., et al. (2013). Optimization of scarless human stem cell genome editing. Nucleic Acids Res, 41(19): p. 9049-61.
79. Perez-Pinera, P., et al. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods, 10(10): p. 973-6.
80. Horlbeck, M.A., et al. (2016). Nucleosomes impede Cas9 access to DNA and. Elife, 5.
81. Knight, S.C., et al. (2015). Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science, 350(6262): p. 823-6.
82. Wu, X., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol, 32(7): p. 670-6.
83. Kim, D., et al. (2016). Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. Genome Res, 26(3): p. 406-15.
84. Kim, D., et al. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat Methods, 12(3): p. 237-43, 1 p following 243.
85. Singh, R., et al. (2015). Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Res, 43(18): p. e118.
86. Gaspar-Maia, A., et al. (2011). Open chromatin in pluripotency and reprogramming. Nat Rev Mol Cell Biol, 12(1): p. 36-47.
87. Tan, E.P., et al. (2015). Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells. Genesis, 53(2): p. 225-36.
88. Veres, A., et al. (2014). Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. Cell Stem Cell, 15(1): p. 27-30.
89. Smith, C., et al. (2014). Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell, 15(1): p. 12-3.
90. Frock, R.L., et al. (2015). Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat Biotechnol, 33(2): p. 179-86.
91. Wang, T., et al. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science, 343(6166): p. 80-4.

92. Fu, Y., et al. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol, 32(3): p. 279-84.
93. Kim, S., et al. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res, 24(6): p. 1012-9.
94. Liang, X., et al. (2015). Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. J Biotechnol, 208: p. 44-53.
95. Mali, P., et al. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol, 31(9): p. 833-8.
96. Ran, F.A., et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell, 154(6): p. 1380-9.
97. Caldecott, K.W. (2008). Single-strand break repair and genetic disease. Nat Rev Genet, 9(8): p. 619-31.
98. Tsai, S.Q., et al. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat Biotechnol, 32(6): p. 569-76.
99. Guilinger, J.P., D.B. Thompson, and D.R. Liu. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat Biotechnol, 32(6): p. 577-82.
100. Wyvekens, N., et al. (2015). Dimeric CRISPR RNA-Guided FokI-dCas9 Nucleases Directed by Truncated gRNAs for Highly Specific Genome Editing. Hum Gene Ther, 26(7): p. 425-31.
101. Shen, B., et al. (2014). Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. Nat Methods, 11(4): p. 399-402.
102. Chiang, T.W., et al. (2016). CRISPR-Cas9(D10A) nickase-based genotypic and phenotypic screening to enhance genome editing. Sci Rep, 6: p. 24356.
103. Makova, K.D. and R.C. Hardison. (2015). The effects of chromatin organization on variation in mutation rates in the genome. Nat Rev Genet, 16(4): p. 213-23.
104. Heyer, W.D., K.T. Ehmsen, and J. Liu. (2010). Regulation of homologous recombination in eukaryotes. Annu Rev Genet, 44(1): p. 113-39.
105. Boward, B., T. Wu, and S. Dalton. (2016). Control of cell fate through cell cycle and pluripotency networks. Stem Cells: p. n/a-n/a.
106. Singh, A.M. and S. Dalton. (2009). The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. Cell Stem Cell, 5(2): p. 141-9.
107. Desmarais, J.A., et al. (2016). Apoptosis and failure of checkpoint kinase 1 activation in human induced pluripotent stem cells under replication stress. Stem Cell Res Ther, 7(1): p. 17.
108. Ciccia, A. and S.J. Elledge. (2010). The DNA damage response: making it safe to play with knives. Mol Cell, 40(2): p. 179-204.
109. Ceccaldi, R., B. Rondinelli, and A.D. D'Andrea. (2016). Repair Pathway Choices and Consequences at the Double-Strand Break. Trends Cell Biol, 26(1): p. 52-64.
110. McVey, M. and S.E. Lee. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet, 24(11): p. 529-38.
111. Dianov, G.L. and U. Hubscher. (2013). Mammalian base excision repair: the forgotten archangel. Nucleic Acids Res, 41(6): p. 3483-90.
112. Song, J., et al. (2016). RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat Commun, 7: p. 10548.
113. Maruyama, T., et al. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat Biotechnol, 33(5): p. 538-42.

114. Mladenov, E., et al. (2016). DNA double-strand-break repair in higher eukaryotes and its role in genomic instability and cancer: Cell cycle and proliferation-dependent regulation. Semin Cancer Biol.
115. Orthwein, A., et al. (2015). A mechanism for the suppression of homologous recombination in G1 cells. Nature, 528(7582): p. 422-6.
116. Mjelle, R., et al. (2015). Cell cycle regulation of human DNA repair and chromatin remodeling genes. DNA Repair (Amst), 30: p. 53-67.
117. Metzger, M.J., et al. (2011). Single-strand nicks induce homologous recombination with less toxicity than double-strand breaks using an AAV vector template. Nucleic Acids Res, 39(3): p. 926-35.
118. Vriend, L.E., et al. (2016). Distinct genetic control of homologous recombination repair of Cas9-induced double-strand breaks, nicks and paired nicks. Nucleic Acids Res.
119. Davis, L. and N. Maizels. (2014). Homology-directed repair of DNA nicks via pathways distinct from canonical double-strand break repair. Proc Natl Acad Sci U S A, 111(10): p. E924-32.
120. Geisinger, J.M., et al. (2016). In vivo blunt-end cloning through CRISPR/Cas9-facilitated non-homologous end-joining. Nucleic Acids Res.
121. Ghezraoui, H., et al. (2014). Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. Mol Cell, 55(6): p. 829-42.
122. Serrano, L., et al. (2011). Homologous recombination conserves DNA sequence integrity throughout the cell cycle in embryonic stem cells. Stem Cells Dev, 20(2): p. 363-74.
123. Tichy, E.D., et al. (2010). Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. Stem Cells Dev, 19(11): p. 1699-711.
124. Cervantes, R.B., et al. (2002). Embryonic stem cells and somatic cells differ in mutation frequency and type. Proc Natl Acad Sci U S A, 99(6): p. 3586-90.
125. Weissbein, U., N. Benvenisty, and U. Ben-David. (2014). Quality control: Genome maintenance in pluripotent stem cells. J Cell Biol, 204(2): p. 153-63.
126. Adams, B.R., et al. (2010). Dynamic dependence on ATR and ATM for double-strand break repair in human embryonic stem cells and neural descendants. PLoS One, 5(4): p. e10001.
127. Fan, J., et al. (2011). Human induced pluripotent cells resemble embryonic stem cells demonstrating enhanced levels of DNA repair and efficacy of nonhomologous end-joining. Mutat Res, 713(1-2): p. 8-17.
128. Liedtke, S., et al. (2015). DNA damage response in neonatal and adult stromal cells compared with induced pluripotent stem cells. Stem Cells Transl Med, 4(6): p. 576-89.
129. Maynard, S., et al. (2008). Human embryonic stem cells have enhanced repair of multiple forms of DNA damage. Stem Cells, 26(9): p. 2266-74.
130. Fung, H. and D.M. Weinstock. (2011). Repair at single targeted DNA double-strand breaks in pluripotent and differentiated human cells. PLoS One, 6(5): p. e20514.
131. Shibata, A., et al. (2011). Factors determining DNA double-strand break repair pathway choice in G2 phase. EMBO J, 30(6): p. 1079-92.
132. Lemaitre, C. and E. Soutoglou. (2014). Double strand break (DSB) repair in heterochromatin and heterochromatin proteins in DSB repair. DNA Repair (Amst), 19: p. 163-8.
133. Goodarzi, A.A. and P.A. Jeggo. (2012). The heterochromatic barrier to DNA double strand break repair: how to get the entry visa. Int J Mol Sci, 13(9): p. 11844-60.
134. Shibata, A. and P.A. Jeggo. (2014). DNA double-strand break repair in a cellular context. Clin Oncol (R Coll Radiol), 26(5): p. 243-9.
135. Dumitru, R., et al. (2012). Human embryonic stem cells have constitutively active Bax at the Golgi and are primed to undergo rapid apoptosis. Mol Cell, 46(5): p. 573-83.

136. Desmarais, J.A., et al. (2012). Human embryonic stem cells fail to activate CHK1 and commit to apoptosis in response to DNA replication stress. Stem Cells, 30(7): p. 1385-93.
137. Barta, T., et al. (2010). Human embryonic stem cells are capable of executing G1/S checkpoint activation. Stem Cells, 28(7): p. 1143-52.
138. Xu, P., et al. (2015). Both TALENs and CRISPR/Cas9 directly target the HBB IVS2–654 (C > T) mutation in β-thalassemia-derived iPSCs. Scientific Reports, 5: p. 12065.
139. Cradick, T.J., et al. (2013). CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity. Nucleic Acids Res, 41(20): p. 9584-92.
140. Sebastiano, V., et al. (2011). In situ genetic correction of the sickle cell anemia mutation in human induced pluripotent stem cells using engineered zinc finger nucleases. Stem Cells, 29(11): p. 1717-26.
141. Zou, J., et al. (2011). Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. Blood, 118(17): p. 4599-4608.
142. Sun, N. and H. Zhao. (2014). Seamless correction of the sickle cell disease mutation of the HBB gene in human induced pluripotent stem cells using TALENs. Biotechnol Bioeng, 111(5): p. 1048-53.
143. Huang, X., et al. (2015). Production of Gene-Corrected Adult Beta Globin Protein in Human Erythrocytes Differentiated from Patient iPSCs After Genome Editing of the Sickle Point Mutation. Stem Cells, 33(5): p. 1470-9.
144. Chang, C.J. and E.E. Bouhassira. (2012). Zinc-finger nuclease-mediated correction of alpha-thalassemia in iPS cells. Blood, 120(19): p. 3906-14.
145. Merkle, F.T., et al. (2015). Efficient CRISPR-Cas9-mediated generation of knockin human pluripotent stem cells lacking undesired mutations at the targeted locus. Cell Rep, 11(6): p. 875-83.
146. Liu, X., et al. (2016). Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. Sci Rep, 6: p. 19675.
147. Smith, C., et al. (2015). Efficient and allele-specific genome editing of disease loci in human iPSCs. Mol Ther, 23(3): p. 570-7.
148. Dias, J., et al. (2011). Generation of red blood cells from human induced pluripotent stem cells. Stem Cells Dev, 20(9): p. 1639-47.
149. Kobari, L., et al. (2012). Human induced pluripotent stem cells can reach complete terminal maturation: in vivo and in vitro evidence in the erythropoietic differentiation model. Haematologica, 97(12): p. 1795-803.
150. Lapillonne, H., et al. (2010). Red blood cell generation from human induced pluripotent stem cells: perspectives for transfusion medicine. Haematologica, 95(10): p. 1651-9.
151. Chang, K.H., et al. (2010). Globin phenotype of erythroid cells derived from human induced pluripotent stem cells. Blood, 115(12): p. 2553-4.
152. Tubsuwan, A., et al. (2013). Parallel assessment of globin lentiviral transfer in induced pluripotent stem cells and adult hematopoietic stem cells derived from the same transplanted beta-thalassemia patient. Stem Cells, 31(9): p. 1785-94.
153. Yang, C.T., et al. (2014). Human induced pluripotent stem cell derived erythroblasts can undergo definitive erythropoiesis and co-express gamma and beta globins. Br J Haematol, 166(3): p. 435-48.
154. Dorn, I., et al. (2015). Erythroid differentiation of human induced pluripotent stem cells is independent of donor cell type of origin. Haematologica, 100(1): p. 32-41.
155. Trakarnsanga, K., et al. (2014). Induction of adult levels of beta-globin in human erythroid cells that intrinsically express embryonic or fetal globin by transduction with KLF1 and BCL11A-XL. Haematologica, 99(11): p. 1677-85.

156. Dominguez, A.A., W.A. Lim, and L.S. Qi. (2016). Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nat Rev Mol Cell Biol, 17(1): p. 5-15.
157. Shalem, O., N.E. Sanjana, and F. Zhang. (2015). High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet, 16(5): p. 299-311.
158. Deng, W., et al. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell, 149(6): p. 1233-1244.
159. Gilbert, L.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell, 154(2): p. 442-51.
160. Balboa, D., et al. (2015). Conditionally Stabilized dCas9 Activator for Controlling Gene Expression in Human Cell Reprogramming and Differentiation. Stem Cell Reports, 5(3): p. 448-59.
161. Maeder, M.L., et al. (2013). Robust, synergistic regulation of human gene expression using TALE activators. Nat Methods, 10(3): p. 243-5.
162. Beerli, R.R., et al. (1998). Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. Proc Natl Acad Sci U S A, 95(25): p. 14628-33.
163. Maeder, M.L., et al. (2013). CRISPR RNA-guided activation of endogenous human genes. Nat Methods, 10(10): p. 977-9.
164. Cheng, A.W., et al. (2013). Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res, 23(10): p. 1163-71.
165. Konermann, S., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature, 517(7536): p. 583-8.
166. Chavez, A., et al. (2015). Highly efficient Cas9-mediated transcriptional programming. Nat Methods, 12(4): p. 326-8.
167. Qi, L.S., et al. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell, 152(5): p. 1173-83.
168. Tanenbaum, M.E., et al. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell, 159(3): p. 635-46.
169. Dahlman, J.E., et al. (2015). Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. Nat Biotechnol, 33(11): p. 1159-61.
170. Kiani, S., et al. (2015). Cas9 gRNA engineering for genome editing, activation and repression. Nat Methods, 12(11): p. 1051-4.
171. Maeder, M.L., et al. (2013). Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nature Biotechnology, 31(12): p. 1137-+.
172. Hilton, I.B., et al. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol, 33(5): p. 510-7.
173. Long, C., et al. (2016). Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy. Science, 351(6271): p. 400-3.
174. Nelson, C.E., et al. (2016). In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. Science, 351(6271): p. 403-7.
175. Tabebordbar, M., et al. (2016). In vivo gene editing in dystrophic mouse muscle and muscle stem cells. Science, 351(6271): p. 407-11.
176. Raya, A., et al. (2010). A protocol describing the genetic correction of somatic human cells and subsequent generation of iPS cells. Nat Protoc, 5(4): p. 647-60.
177. Ran, F.A., et al. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc, 8(11): p. 2281-308.
178. Yusa, K., et al. (2011). Targeted gene correction of alpha1-antitrypsin deficiency in induced pluripotent stem cells. Nature, 478(7369): p. 391-4.

179. Kim, T.K. and J.H. Eberwine. (2010). Mammalian cell transfection: the present and the future. Anal Bioanal Chem, 397(8): p. 3173-8.
180. Hendriks, W.T., et al. (2015). TALEN- and CRISPR/Cas9-Mediated Gene Editing in Human Pluripotent Stem Cells Using Lipid-Based Transfection. Curr Protoc Stem Cell Biol, 34: p. 5B 3 1-25.
181. Hendriks, W.T., C.R. Warren, and C.A. Cowan. (2016). Genome Editing in Human Pluripotent Stem Cells: Approaches, Pitfalls, and Solutions. Cell Stem Cell, 18(1): p. 53-65.
182. Yusa, K. (2013). Seamless genome editing in human pluripotent stem cells using custom endonuclease–based gene targeting and the piggyBac transposon. Nat. Protocols, 8(10): p. 2061-2078.
183. Zhu, X., et al. (2014). An efficient genotyping method for genome-modified animals and human cells generated with CRISPR/Cas9 system. Sci Rep, 4: p. 6420.
184. Yusa, K., et al. (2011). A hyperactive piggyBac transposase for mammalian applications. Proc Natl Acad Sci U S A, 108(4): p. 1531-6.
185. Zhou, H., et al. (2009). Generation of induced pluripotent stem cells using recombinant proteins. Cell Stem Cell, 4(5): p. 381-4.
186. Schlaeger, T.M., et al. (2015). A comparison of non-integrating reprogramming methods. Nat Biotechnol, 33(1): p. 58-63.
187. Wang, Y., et al. (2009). Generation of induced pluripotent stem cells from human beta-thalassemia fibroblast cells. Cell Res, 19(9): p. 1120-3.
188. Treisman, R., S.H. Orkin, and T. Maniatis. (1983). Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. Nature, 302(5909): p. 591-6.
189. Westaway, D. and R. Williamson. (1981). An intron nucleotide sequence variant in a cloned beta +-thalassaemia globin gene. Nucleic Acids Res, 9(8): p. 1777-88.
190. Ovchinnikov, D.A., J. Sun, and E.J. Wolvetang. (2015). Generation of Footprint-Free Induced Pluripotent Stem Cells from Human Fibroblasts Using Episomal Plasmid Vectors. Methods Mol Biol, 1330: p. 37-45.
191. Yu, J., et al. (2011). Efficient feeder-free episomal reprogramming with small molecules. PLoS One, 6(3).
192. Carey, B.W., et al. (2011). Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. Cell Stem Cell, 9(6): p. 588-598.
193. Soldner, F., et al. (2009). Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell, 136(5): p. 964-77.
194. Hotta, A. and S. Yamanaka. (2015). From Genomics to Gene Therapy: Induced Pluripotent Stem Cells Meet Genome Editing. Annu Rev Genet, 49: p. 47-70.
195. Wirt, S.E. and M.H. Porteus. (2012). Development of nuclease-mediated site-specific genome modification. Curr Opin Immunol, 24(5): p. 609-16.
196. Inui, M., et al. (2014). Rapid generation of mouse models with defined point mutations by the CRISPR/Cas9 system. Sci Rep, 4: p. 5396.
197. Kass, E.M. and M. Jasin. (2010). Collaboration and competition between DNA double-strand break repair pathways. FEBS Lett, 584(17): p. 3703-8.
198. Yang, H., H.Y. Wang, and R. Jaenisch. (2014). Generating genetically modified mice using CRISPR/Cas-mediated genome engineering. Nature Protocols, 9(8): p. 1956-1968.
199. Vouillot, L., A. Thelie, and N. Pollet. (2015). Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. G3 (Bethesda), 5(3): p. 407-15.
200. Qiu, P., et al. (2004). Mutation detection using Surveyor nuclease. Biotechniques, 36(4): p. 702-7.

201. Chatterjee, P., Y. Cheung, and C. Liew. (2011). Transfecting and nucleofecting human induced pluripotent stem cells. J Vis Exp, (56): p. e3110.
202. Soldner, F., et al. (2011). Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. Cell, 146(2): p. 318-31.
203. Miyaoka, Y., et al. (2014). Isolation of single-base genome-edited human iPS cells without antibiotic selection. Nat Methods, 11(3): p. 291-3.
204. Chu, V.T., et al. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol, 33(5): p. 543-8.
205. Orlando, S.J., et al. (2010). Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. Nucleic Acids Res, 38(15): p. e152.
206. Bae, S., et al. (2014). Microhomology-based choice of Cas9 nuclease target sites. Nat Methods, 11(7): p. 705-6.
207. Kim, Y., J. Kweon, and J.S. Kim. (2013). TALENs and ZFNs are associated with different mutation signatures. Nat Methods, 10(3): p. 185.
208. Certo, M.T., et al. (2011). Tracking genome engineering outcome at individual DNA breakpoints. Nat Methods, 8(8): p. 671-6.
209. Ramirez, C.L., et al. (2012). Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. Nucleic Acids Res, 40(12): p. 5560-8.
210. Moreno-Mateos, M.A., et al. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods, 12(10): p. 982-8.
211. McCormick, M. (1987). Sib selection. Methods Enzymol, 151: p. 445-9.
212. Flynn, R., et al. (2015). CRISPR-mediated genotypic and phenotypic correction of a chronic granulomatous disease mutation in human iPS cells. Exp Hematol, 43(10): p. 838-848 e3.
213. Hall, D.B. and K. Struhl. (2002). The VP16 activation domain interacts with multiple transcriptional components as determined by protein-protein cross-linking in vivo. J Biol Chem, 277(48): p. 46043-50.
214. Wang, L., S.R. Grossman, and E. Kieff. (2000). Epstein-Barr virus nuclear protein 2 interacts with p300, CBP, and PCAF histone acetyltransferases in activation of the LMP1 promoter. Proc Natl Acad Sci U S A, 97(1): p. 430-5.
215. Kearns, N.A., et al. (2014). Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. Development, 141(1): p. 219-23.
216. Kabadi, A.M., et al. (2014). Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Res, 42(19): p. e147.
217. Yamano, S., J. Dai, and A.M. Moursi. (2010). Comparison of transfection efficiency of nonviral gene transfer reagents. Mol Biotechnol, 46(3): p. 287-300.
218. Gao, X., et al. (2014). Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers. Nucleic Acids Res, 42(20): p. e155.
219. Manoj, P. (2016). Droplet digital PCR technology promises new applications and research areas. Mitochondrial DNA, 27(1): p. 742-6.
220. Maier, T., M. Guell, and L. Serrano. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett, 583(24): p. 3966-73.
221. Goh, P.A., et al. (2013). A systematic evaluation of integration free reprogramming methods for deriving clinically relevant patient specific induced pluripotent stem (iPS) cells. PLoS One, 8(11): p. e81622.
222. Aloni-Grinstein, R., et al. (2014). p53: the barrier to cancer stem cell formation. FEBS Lett, 588(16): p. 2580-9.
223. Dang, C.V., et al. (1999). Function of the c-Myc oncogenic transcription factor. Exp Cell Res, 253(1): p. 63-77.

224. Takahashi, K. and S. Yamanaka. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. Nat Rev Mol Cell Biol, 17(3): p. 183-93.
225. Meng, X.M., et al. (2012). Efficient Reprogramming of Human Cord Blood CD34(+) Cells Into Induced Pluripotent Stem Cells With OCT4 and SOX2 Alone. Molecular Therapy, 20(2): p. 408-416.
226. Buganim, Y., et al. (2014). The developmental potential of iPSCs is greatly influenced by reprogramming factor selection. Cell Stem Cell, 15(3): p. 295-309.
227. Robinton, D.A. and G.Q. Daley. (2012). The promise of induced pluripotent stem cells in research and therapy. Nature, 481(7381): p. 295-305.
228. Musunuru, K. (2013). Genome editing of human pluripotent stem cells to generate human cellular disease models. Dis Model Mech, 6(4): p. 896-904.
229. Yang, J., et al. (2016). Induced pluripotent stem cells in Alzheimer's disease: applications for disease modeling and cell-replacement therapy. Mol Neurodegener, 11(1): p. 39.
230. Doench, J.G., et al. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol, 32(12): p. 1262-7.
231. Xu, H., et al. (2015). Sequence determinants of improved CRISPR sgRNA design. Genome Res, 25(8): p. 1147-57.
232. Gafni, O., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. Nature, 504(7479): p. 282-6.
233. Byrne, S.M., et al. (2015). Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells. Nucleic Acids Res, 43(3): p. e21.
234. Hoban, M.D. and D.E. Bauer. (2016). A genome editing primer for the hematologist. Blood, 127(21): p. 2525-35.
235. Canver, M.C., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature, 527(7577): p. 192-7.
236. Vierstra, J., et al. (2015). Functional footprinting of regulatory DNA. Nat Methods, 12(10): p. 927-30.
237. Halford, S.E., et al. (2011). The reaction mechanism of FokI excludes the possibility of targeting zinc finger nucleases to unique DNA sites. Biochem Soc Trans, 39(2): p. 584-8.
238. Gao, F., et al. (2016). DNA-guided genome editing using the Natronobacterium gregoryi Argonaute. Nat Biotech, advance online publication.
239. Bauer, D.E., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science, 342(6155): p. 253-7.